**Xxx课程大作业报告**

**茹悦**

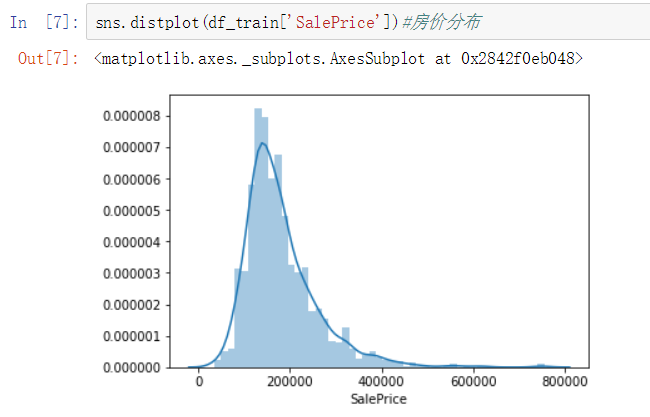
1. 总体介绍

本次大作业选取kaggle上房价数据集，在做了一些房价与其他因素的相关性分析后，选出一些属性来打造房价预测模型，最终达到85+%的准确度。

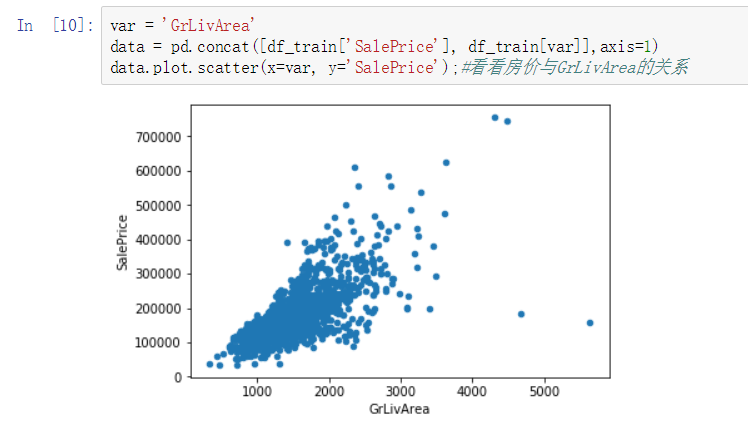
1. 数据集分析与可视化(用了seaborn库，包装了一些画图函数比matplotlib好用些)



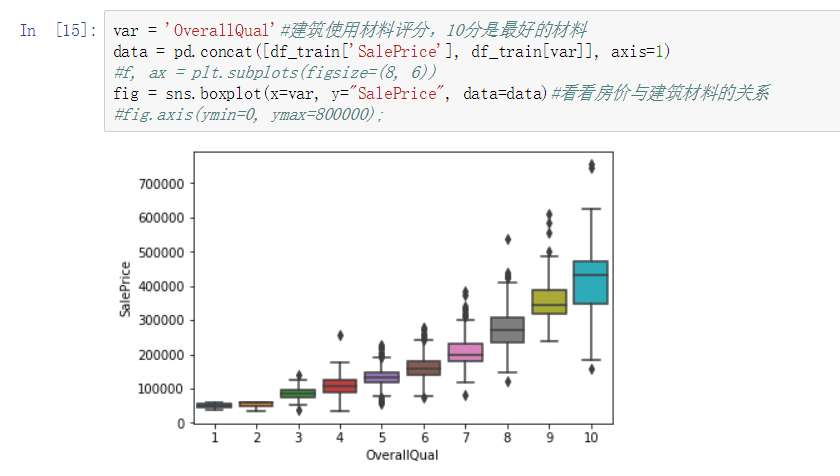
首先读入train.csv文件内容到dataframe中，查看有哪些属性



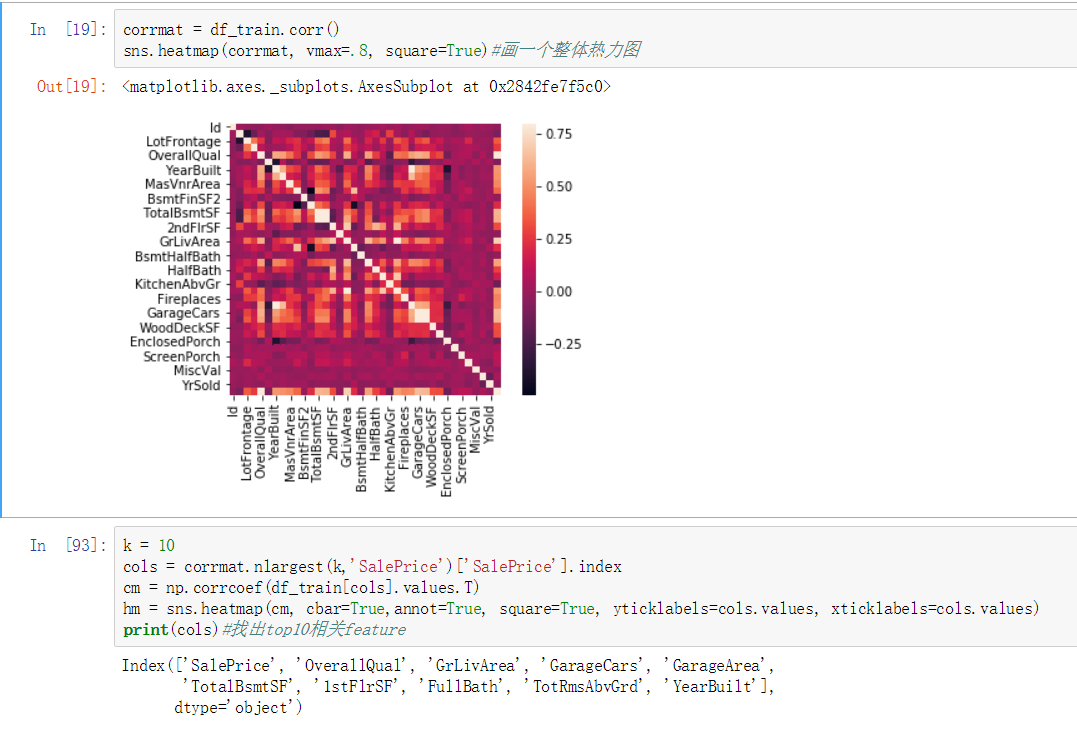
首先看以下房价分布的情况，这与我们预期一样，大部分集中在10-20万美元区间，豪宅数目占比很低



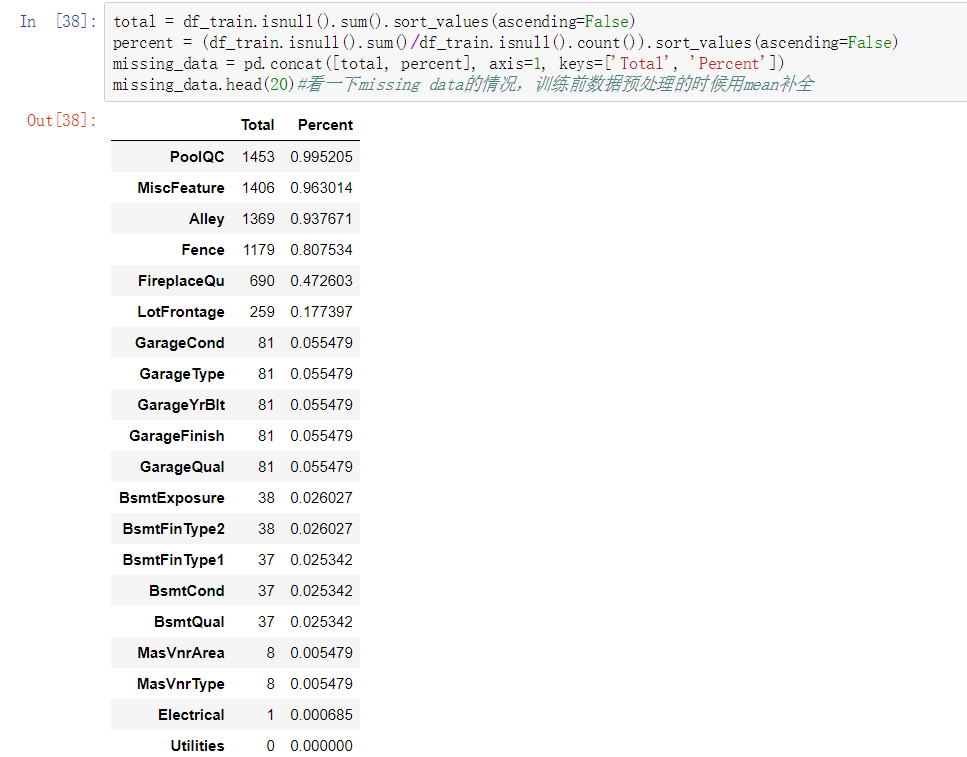
看一下房价与房子建筑面积之间的关系



再看一下建筑材料和房价之间的关系，也是正相关的

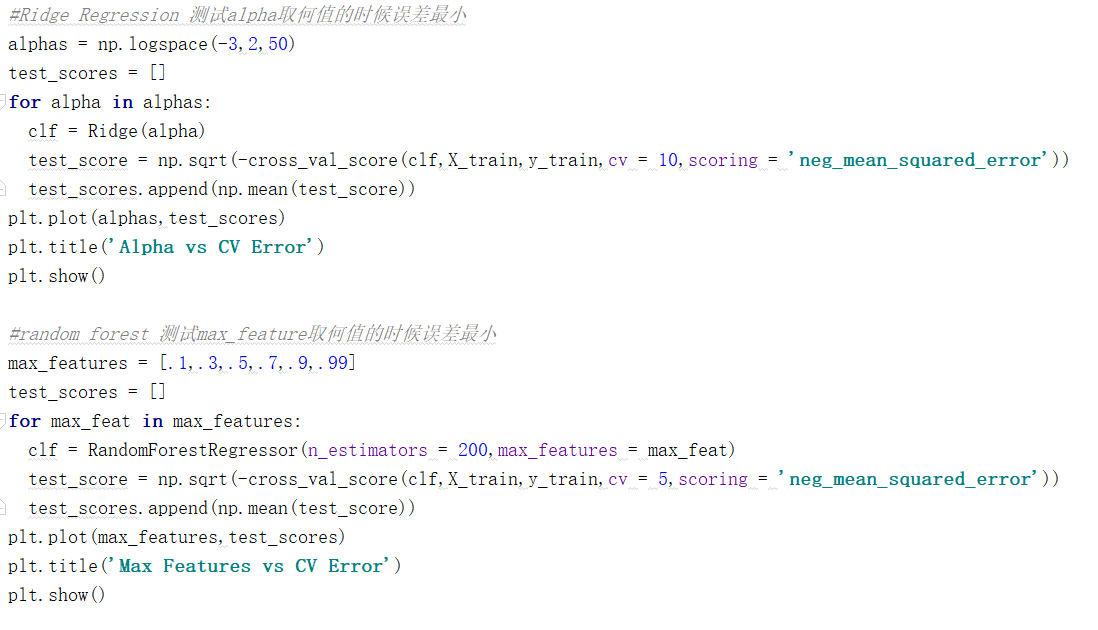


再画了一个热力图，找出前十个正相关系数最大的，之后用相关系数最大的几个feature来搭建模型

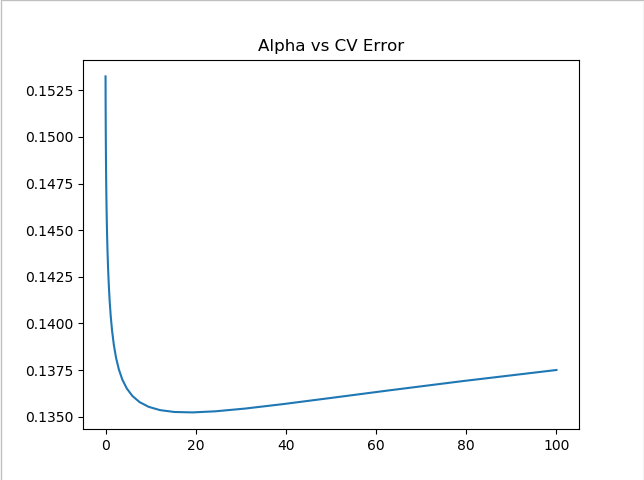


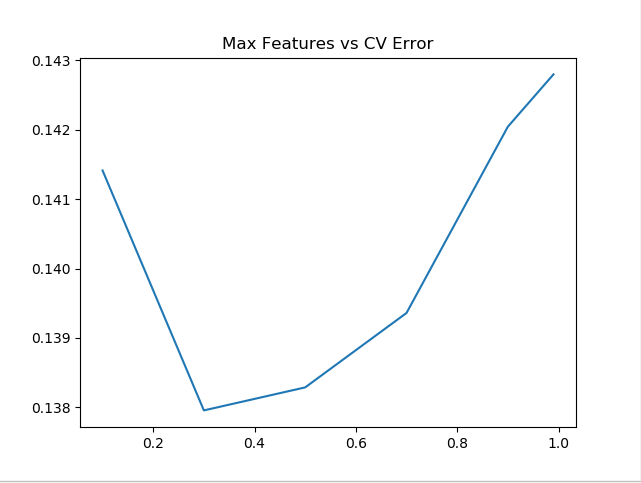
最后看一下数据缺失的情况，训练模型之前要用均值填充缺失的数据。

训练预测模型中数据预处理阶段不再赘述，主要是填充缺失数据，生成训练和验证集，这里我用了一种回归模型和一个随机森林模型，并将两者结合在一起形成最终的模型。

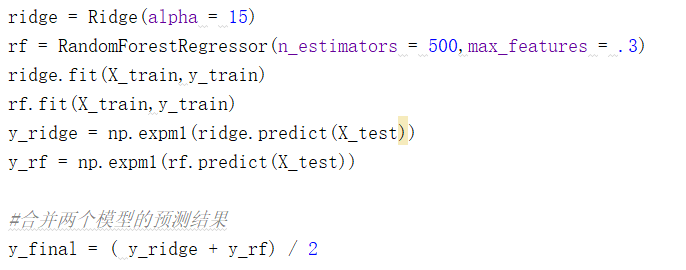


两个模型的超参数遍历一系列值，找到最优情况：





可以看到随机森林模型max\_features参数最好设为0.3，回归模型中alpha参数最好设为15左右，都能达到约87%的准确度



为了简化复杂度，这里直接将两个模型的权重都设为0.5

1. 总结

本次大作业学习了seaborn数据可视化库的一些简单用法，并学习使用sklearn机器学习库的一些简单的机器学习模型使用，缺点是没有了解这些机器学习算法背后的原理而只是调用提供的接口而已，也没有使用深度学习相关的算法。