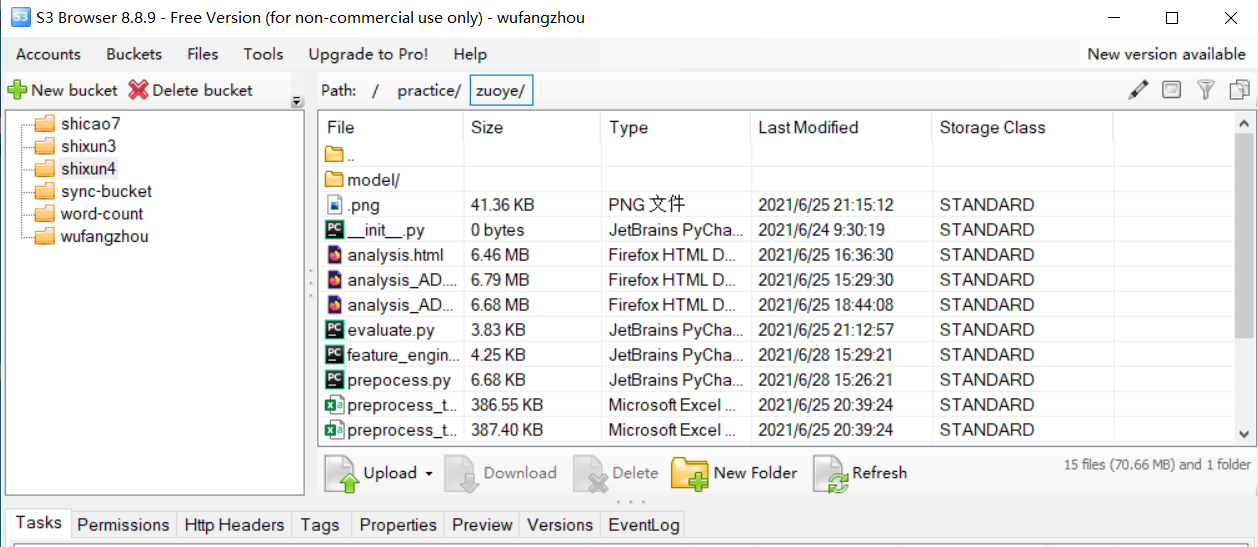
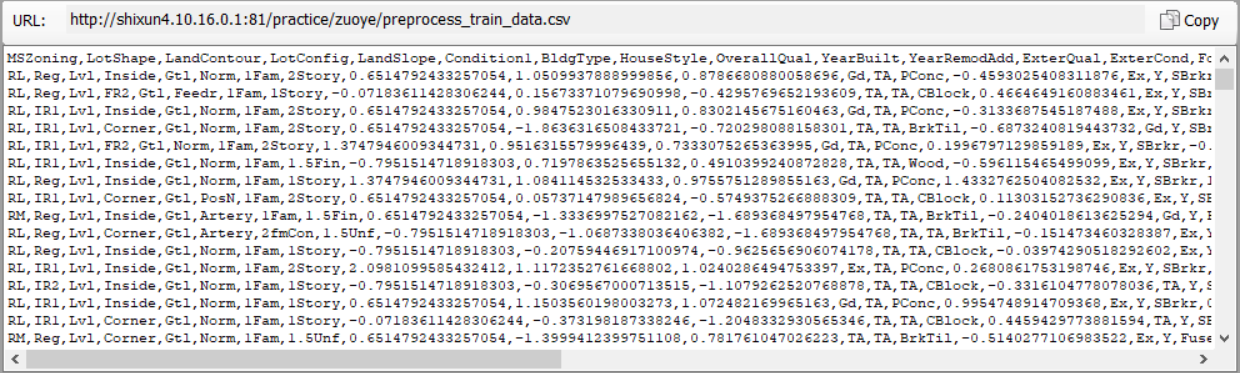
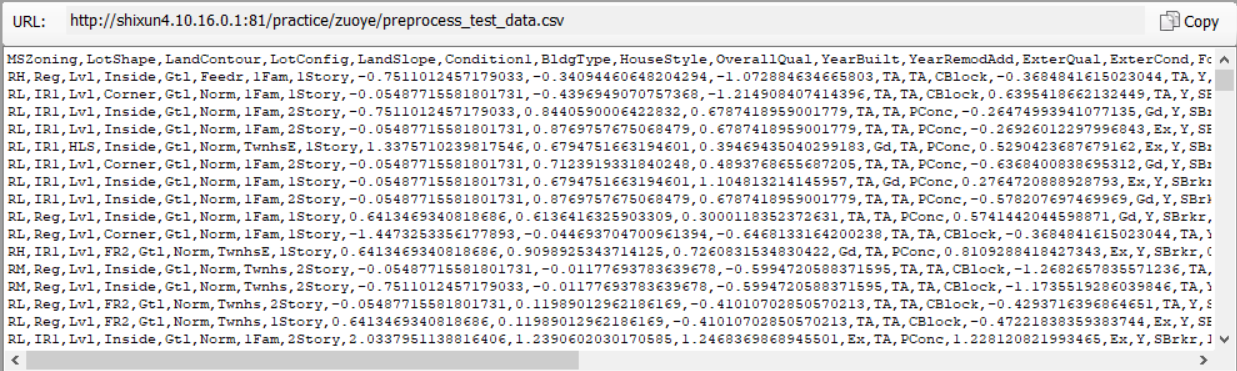
设计说明书及流程

我所选择的数据库是kaggle的房价预测数据库，并将房价离散化，转变为分类任务，代码、数据等同步到s3的shixun4 bucket下的practice/zuoye目录

一、运行preproces.py

进行数据分析，输出基本信息、检讨分析、关联图并生成详细分析报告analysis.html。之后将其作为依照进行缺失值处理，如去掉信息重复的特征和缺失值过多的特征并用众数补全剩下的缺失值。再继续把数据分析结果作为依照去除关联性较低且信息重复的无意义的特征，如id等。之后将特征和房价标准化和离散化。生成preprocess\_train\_data.csv、preprocess\_test\_data.csv和result.csv

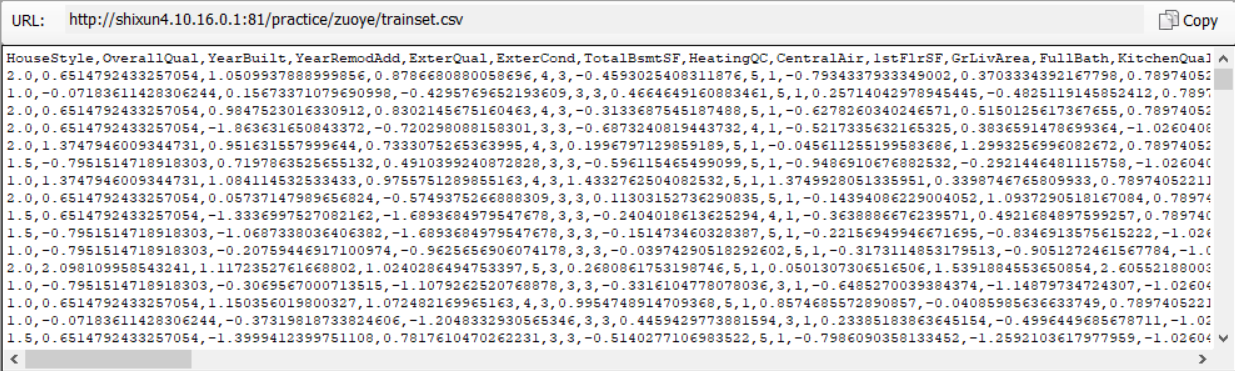


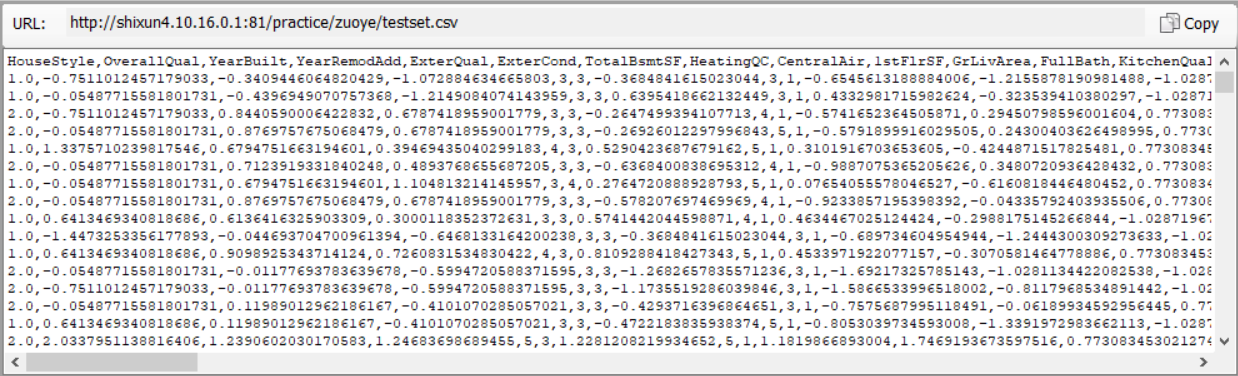




二、运行feature\_engineering.py

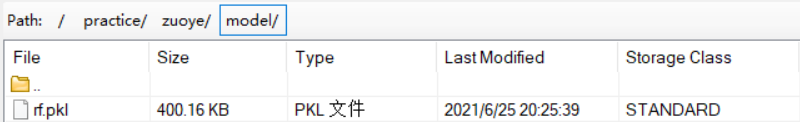
在这里处理非数值离散特征，进行独热编码或赋予离散值。生成trainset.csv和testset.csv





三、运行rf.py

在目录model下生成随机森林模型rf.pkl



四、运行evaluate.py

用模型预测测试集结果，并进行评估，输出准确率、精确率等并生成各种可视化报告。

