设计说明书

我使用的是kaggle上的房价预测的数据集。

设计的主要目标是

1. 观察每个变量特征的意义以及对于问题的重要程度
2. 研究因变量的特征：就是房价
3. 研究其他多变量对房价的影响以及他们之间的关系
4. 对一些缺失数据，异常点和分类数据进行处理
5. 训练，验证，评估模型

主要流程

1. 读取观察数据集
2. 观察各项主要特征与房价的关系
3. 提取主要特征
4. 验证主要特征是否满足要求（通过相关系数矩阵验证）
5. 训练模型，挑选了svm，随机森林和贝叶斯线性回归来训练，根据损失函数大致挑选最适合的
6. 对数据进行清洗
7. 使用随机森林模型训练，验证，评估