2022-2023学年第一学期课程

（非试卷）

**课程代码：05b00069**

**适用年级：2019**

**适用专业：数据科学与大数据技术**

考核要求：

台风灾害期间房屋损坏率预测。该数据集来源于荷兰红十字会提供的510全球数据库，包括过去二十年来菲律宾发生的12次典型台风的数据，数据见文件all.csv。以下是这些台风的名称：“Bopha”,“Goni”, “Hagupit”, “Haima”, “Haiyan”, Kalmaegi”, “Koppu”, “Melor”, “Nock-Ten”, “Rammasun”, “Sarika”和“Utor”。台风灾害数据包括1638次观测值。

下面简介预测目标、台风事件相关变量、现场地形特征和城市社会经济指标相关的变量。其他部分变量也可由于英文单词可推出它的含义。

**预测目标:** 房屋损坏率的变量。

Total.damaged.houses..rel..：房屋损坏率

**危险指标:**与台风事件相关的变量。

Windspeed [km/h]:风速

Distance.to.typhoon [km]:台风路径到城市中心的最短距离

Rainfall [mm]:降雨量

Distance\_first\_impact [km]:从台风登陆点到城市中心的距离

Experience[-]:本次台风前，该市所属地区承受台风的平均次数

**现场和社会经济指标：**与现场地形特征和城市社会经济指标相关的变量。只取决于市政当局，而不取决于台风（假设它们在时间上相对稳定）。

Elevation [m]:城市的平均海拔

Slope [°]:城市的平均地形坡度

Slope\_stdev [-]:城市地形坡度的标准差

Ruggedness [-]:城市的平均地形崎岖度

Ruggedness\_stdev [-]:城市地形崎岖度的标准偏差

Population.density [hab/km2]:人口密度

Poverty.incidence [-]:贫困发生率

X..skilled.Agriculture.Forestry.Fishermen [-]:熟练的农业/林业/渔民的比例

X..strong.roof.type [-]:强屋顶类型比例

X..strong.wall.type [-]:强墙类型比例

要求：

1. 数据探索分析
2. 数据清洗，比如异常值、缺失值、数据标准化等
3. 特征选择（可用不同特征选择方法）
4. 模型训练与预测（可用不同算法），注：20%为测试集，80%为训练集
5. 模型评估与比较（可用多个评估算法比较）

**提交说明：**

1）新建一个文件夹，命名为座号+姓名，如“01\_陈小明”，同时作品程序中修改当前目录如“D:/01\_陈小明” ，文件夹内分别两个子文件夹（命名为“code”、“data”）和一个本设计文档（命名如“陈小明\_设计文档.docx”）。“code”、“data”文件夹分别存放代码文件和数据，不得有冗余无用文件。

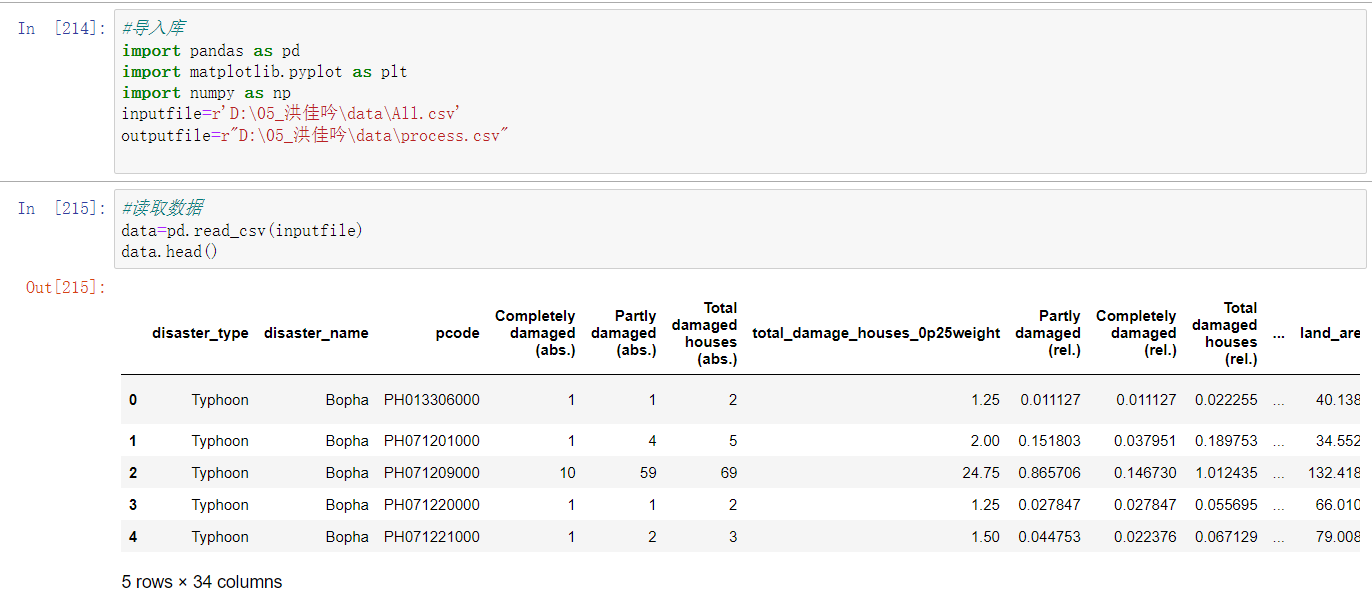
2）设计文档按步骤完成，必须有说明，代码、必要注释，并附上每个步骤运行结果。代码思路清晰、代码可执行、交互友好、注释清楚。文件不冗余，命名要规范。

3）整个文件夹打包后提交平台，同时设计文档文件须打印出来（双面打印）。

答题部分

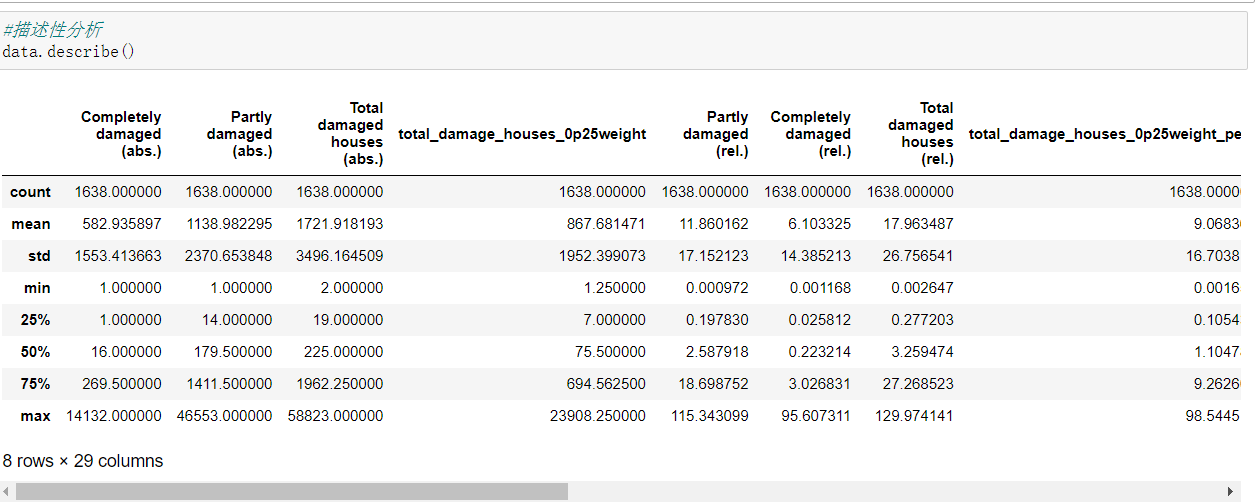
1. 数据探索分析
2. 导入数据

通过预览前5行数据可见，该数据有34列，即有34个属性。

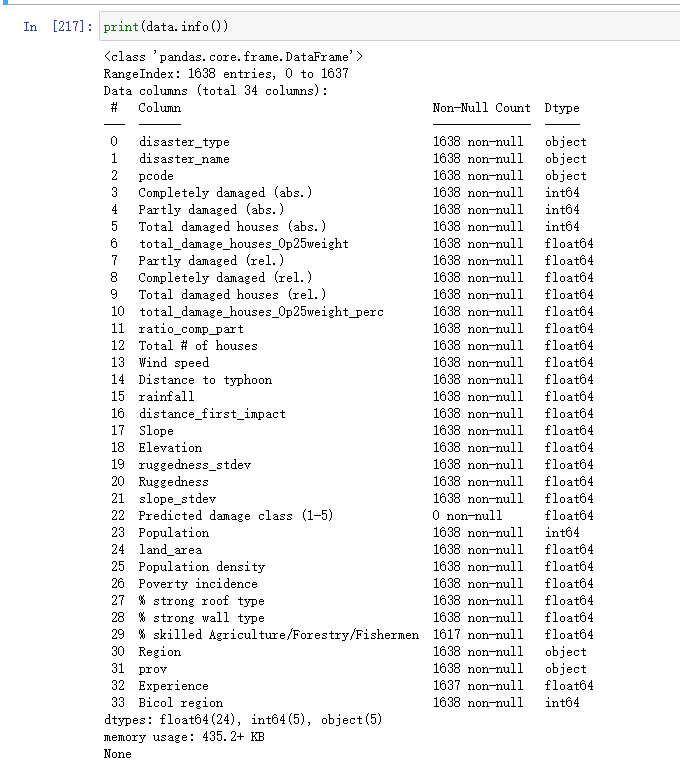


1. 描述性分析

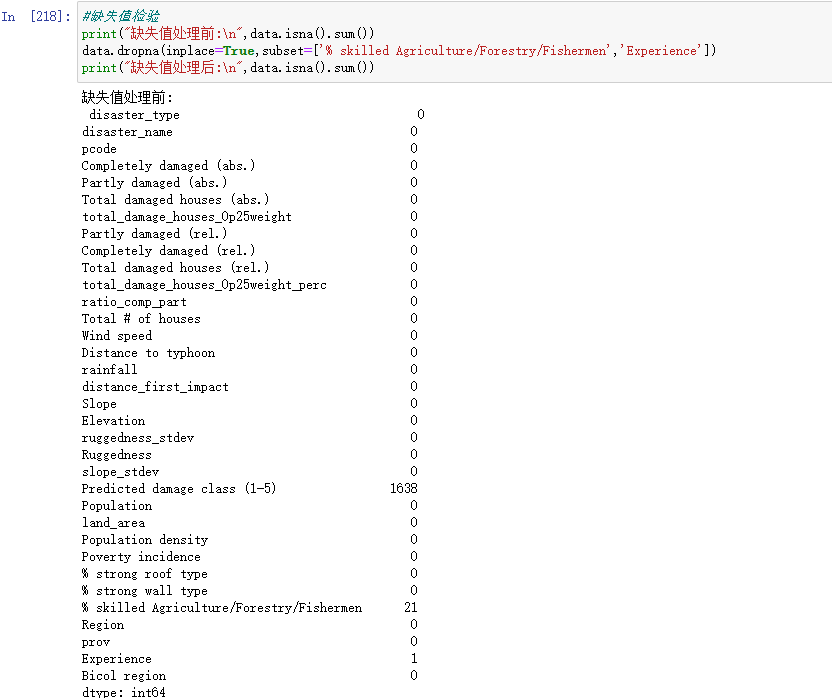
通过describe()可见部分列数据的count索引不为1638，可见存在缺失值。

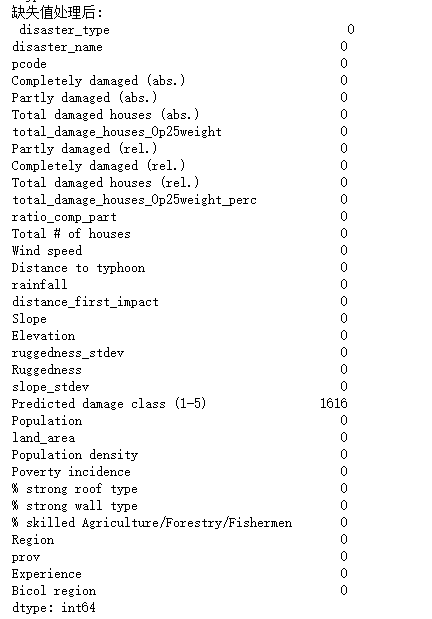


通过info()可见，预测变量数据为float型，目标变量为int型，类别特征为object型，这里类别属性不用作训练数据，故去除。



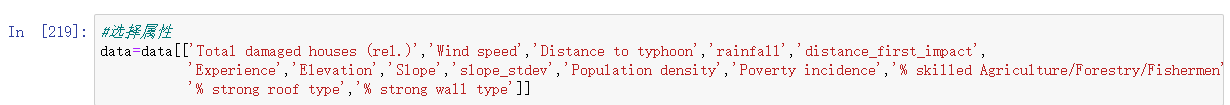
1. 数据清洗，比如异常值、缺失值、数据标准化等
2. 缺失值检验



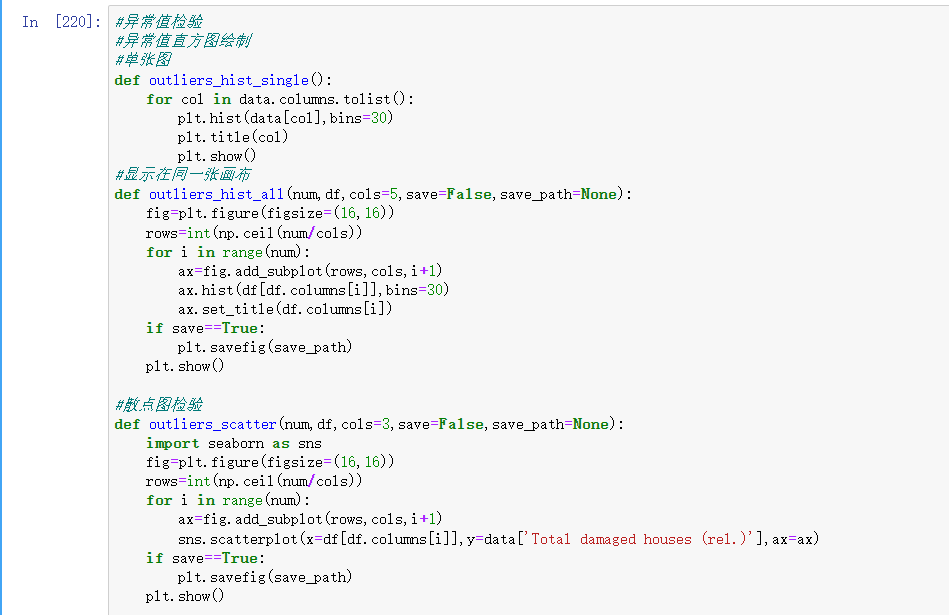


1. 属性选择

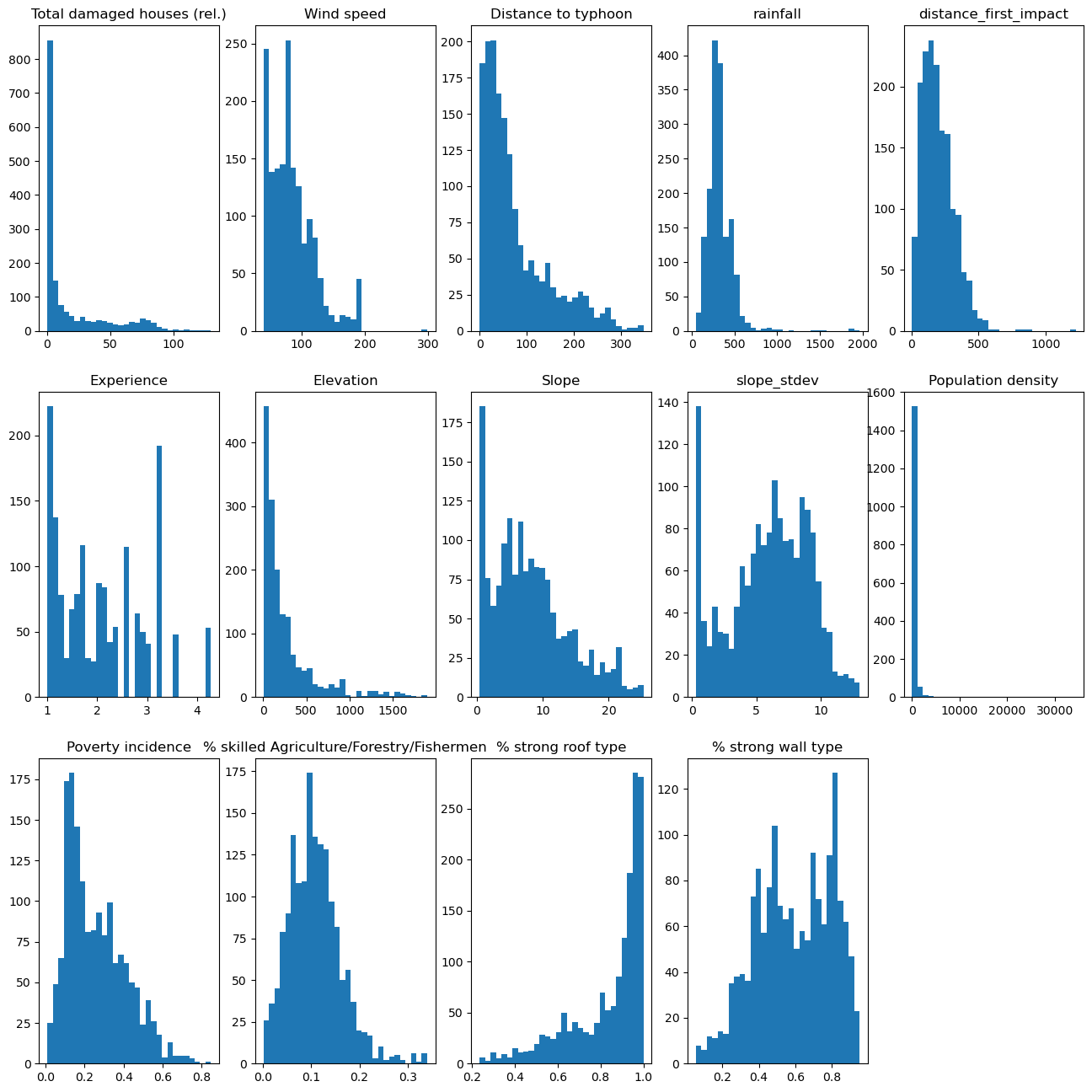
根据数据探索分析，此处选择有用属性，剔除无用属性。

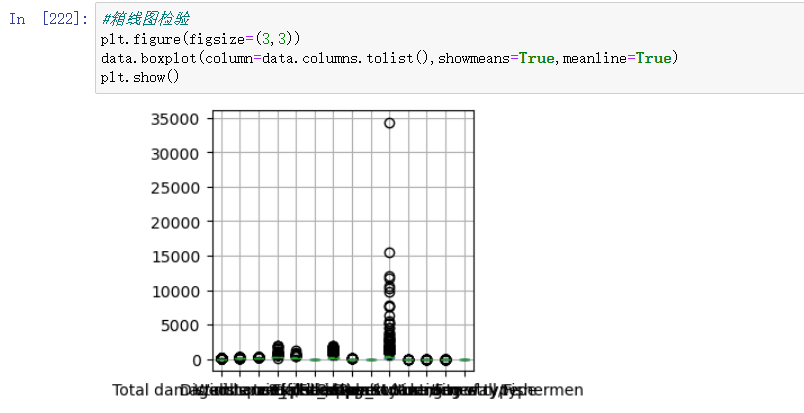


1. 异常值检验

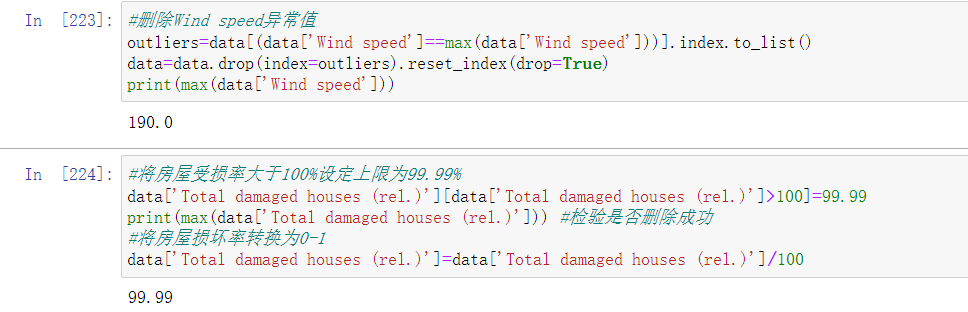


根据下面直方图和箱型图检验结果可见，该数据呈现偏态分布，不服从正态分布，对后续回归分析会产生影响，且Wind Speed存在离群点，Total damaged houses(rel.)存在超出100，不符合常规的数据，故将超过100的值设定为上限99.99。



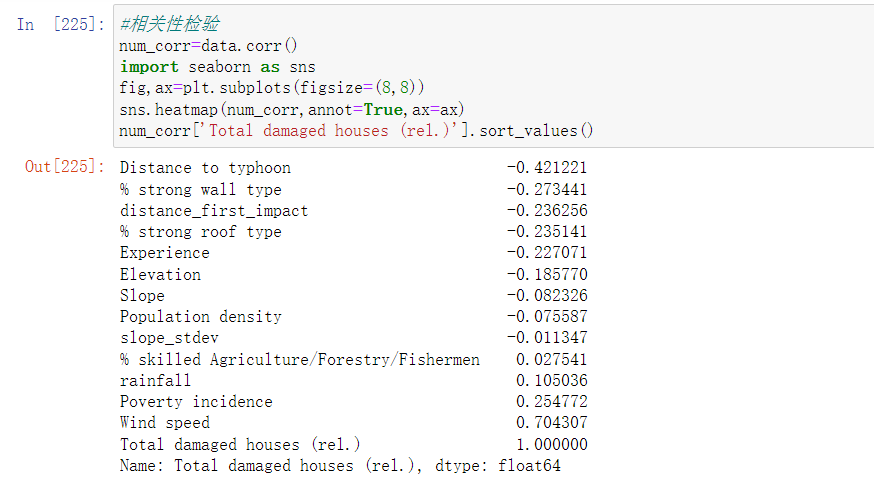


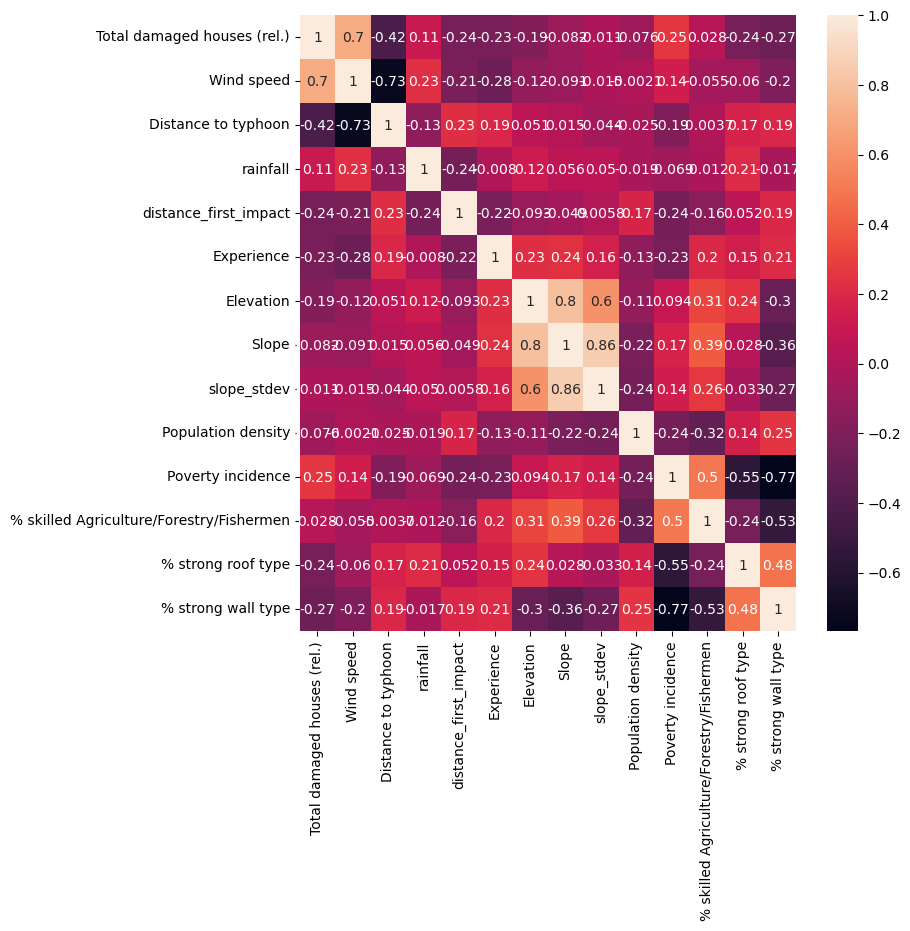
1. 剔除异常值



1. 相关性检验，查看数值特征之间的相关关系

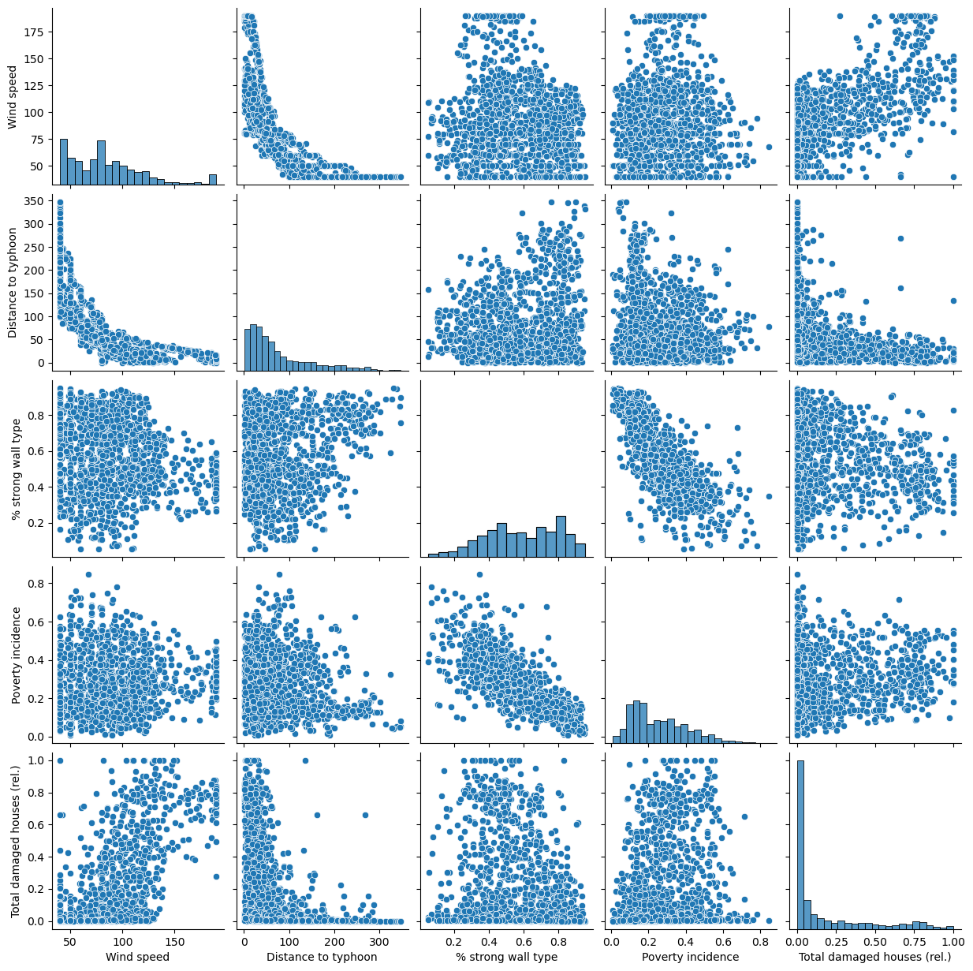
通过计算相关系数可以得到与目标变量的相关程度



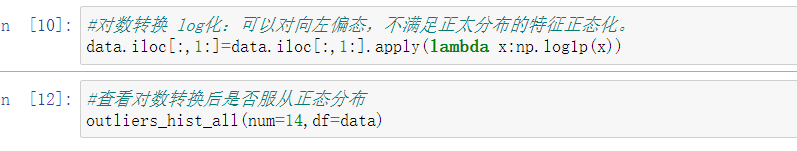


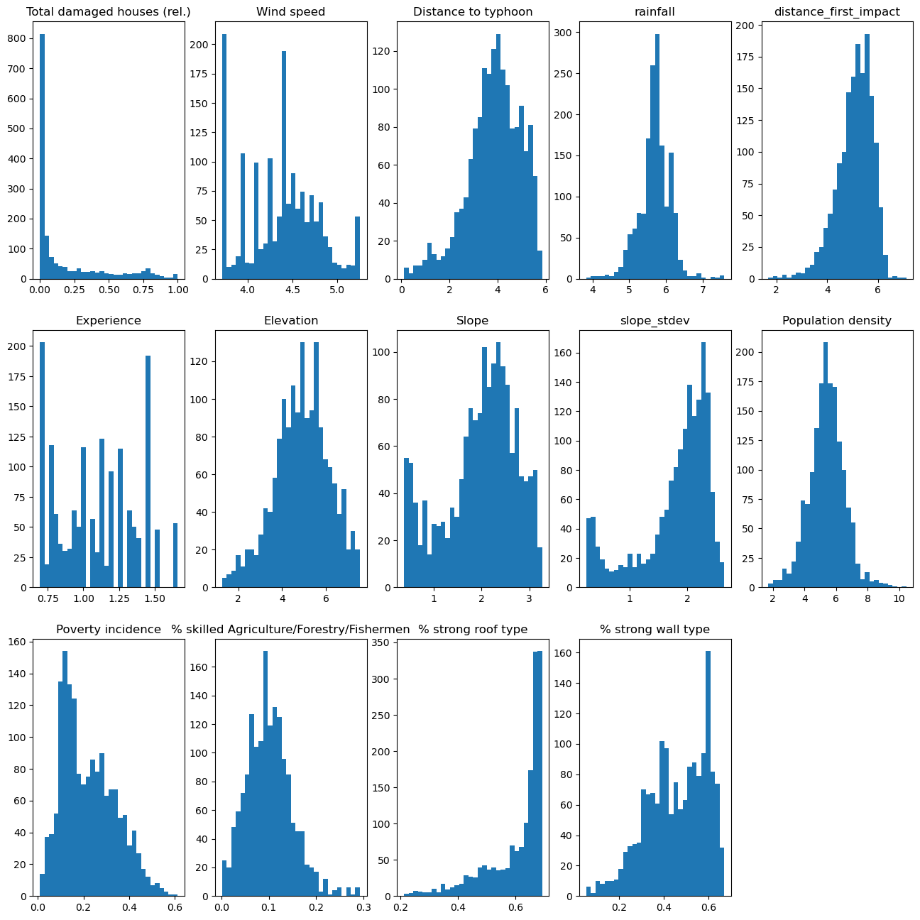
1. 多变量研究

通过（5）得到的前四个与预测变量相关性最大的特征进行展示特征之间关系，可见变量之间存在明显的相关性。

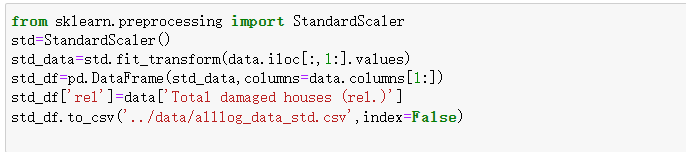


1. 根据(3)的直方图进行数据转换，使其正态化。





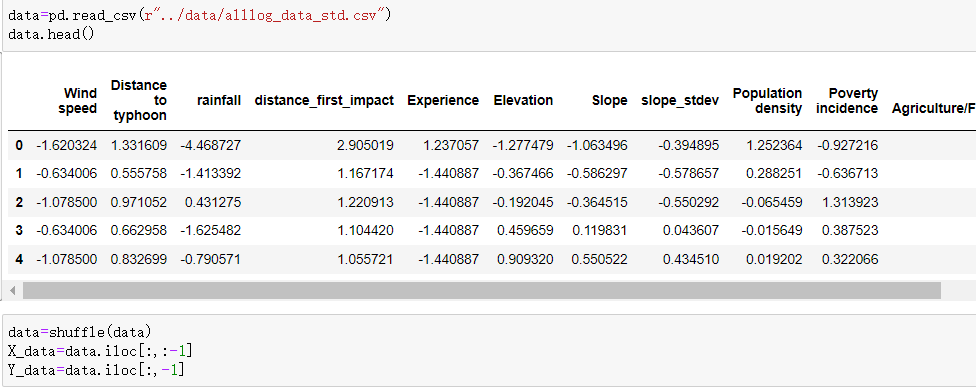
1. 将特征数据标准化，并保存为alllog\_data\_std.csv。



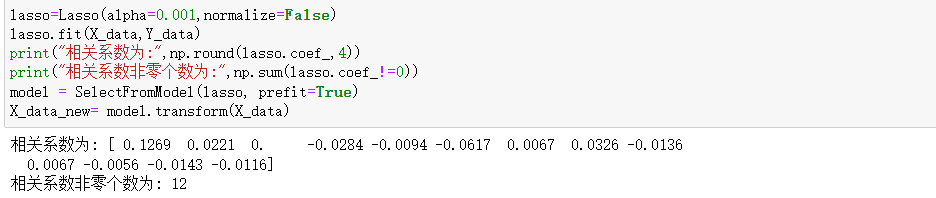
1. 特征选择
2. 导入所需的库



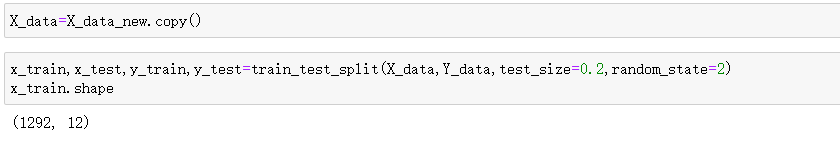
1. 读取数据，打乱数据并分离目标变量与预测变量。

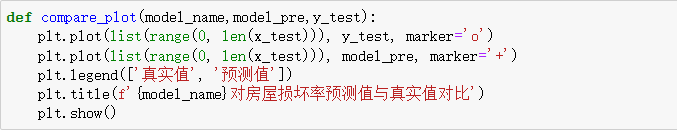


1. LASSO降维

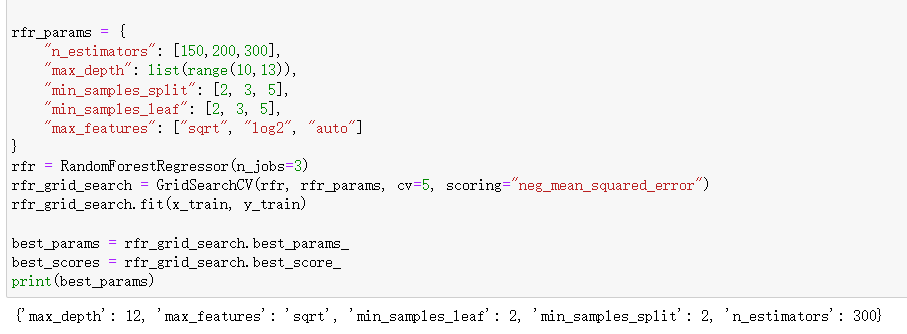


1. 模型训练与预测（20%为测试集，80%为训练集）。
2. 分割数据集及定义真实值与预测值折现对比图函数



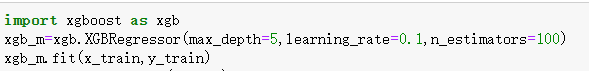


1. 网格搜索构建随机森林回归模型

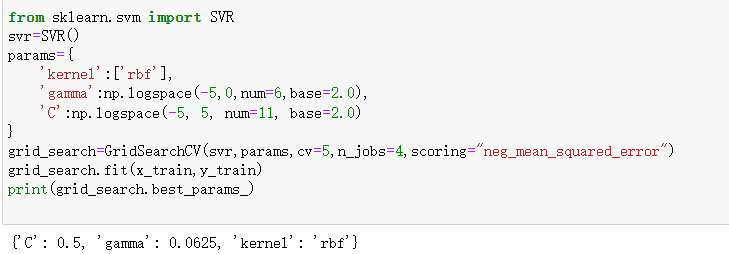




1. 构建xgb模型

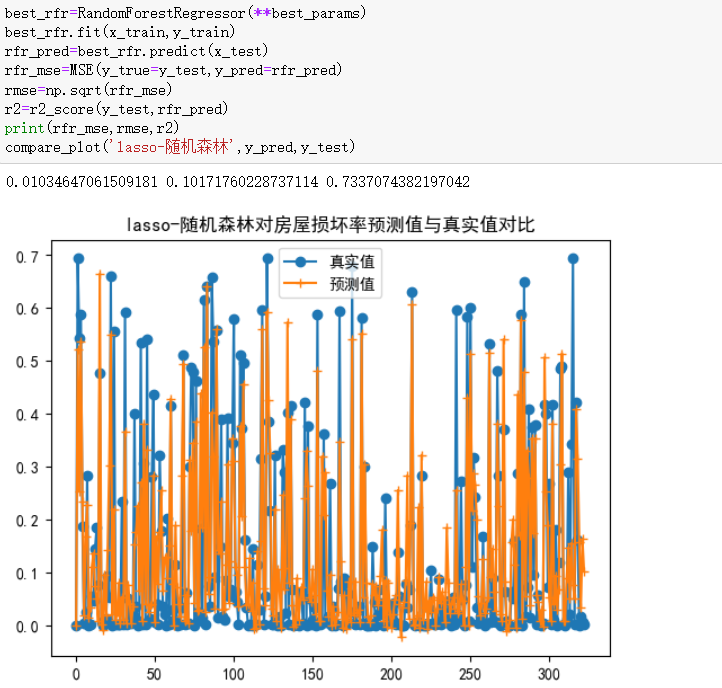


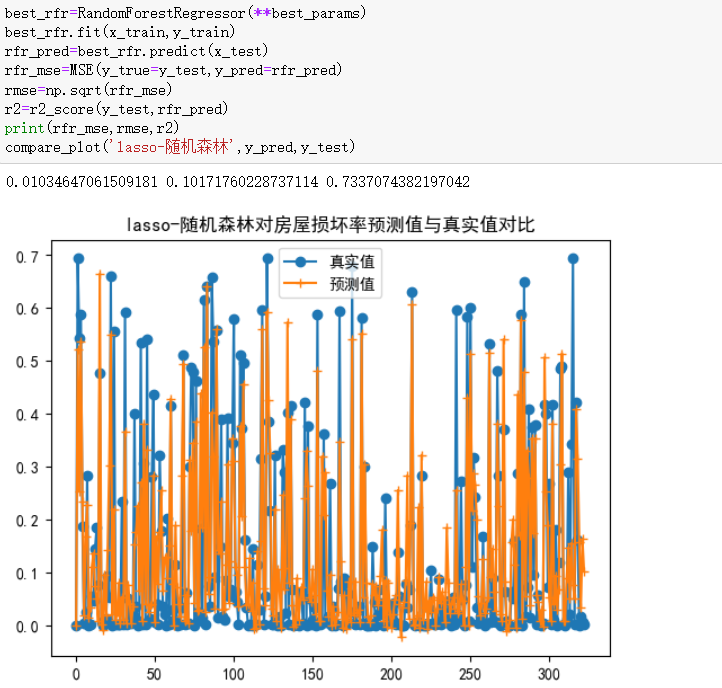
1. 构建SVR模型



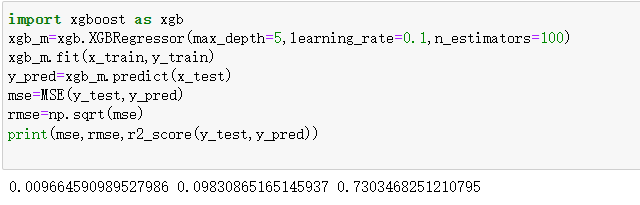


1. 模型评估与比较
2. LASSO-随机森林

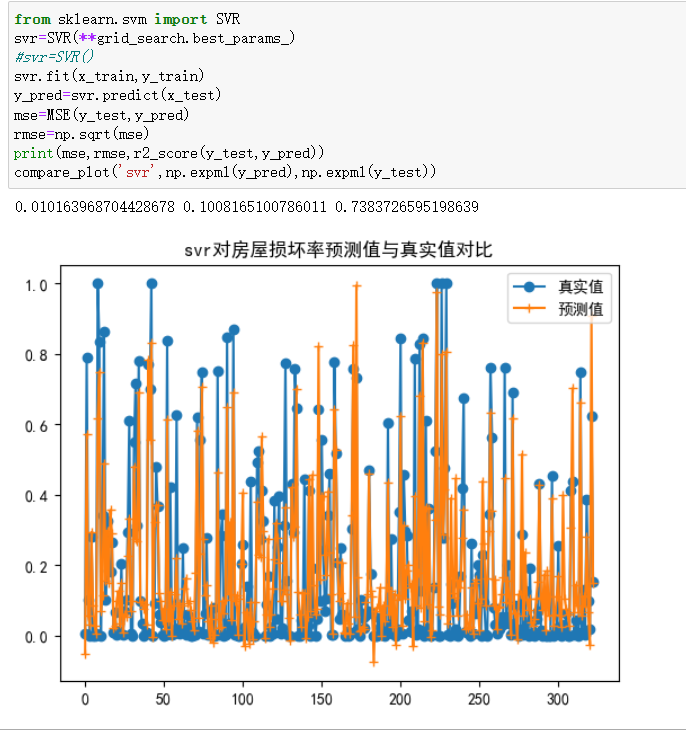


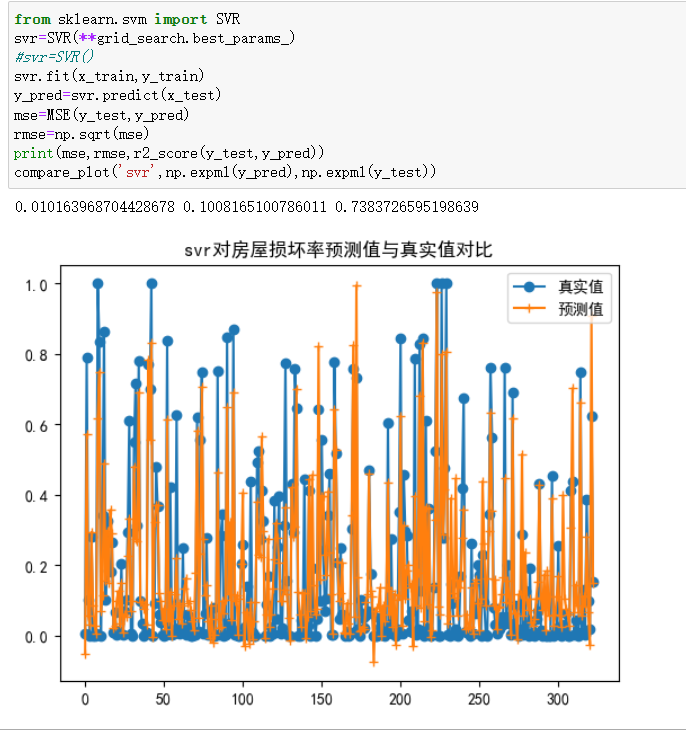


1. LASSO-XGB



1. LASSO-SVR





通过上述三个模型的运行结果可见，在使用正则化参数为0.001的Lasso降维后训练出来的三个模型，评估指标相差较小，都在73%上下浮动，拟合效果相差较小，均方根误差都在0.1左右，其中rmse和mse最小、r2最大的是LASSO-SVR。