# 大数据人工智能大作业使用说明书

## 作业环境：

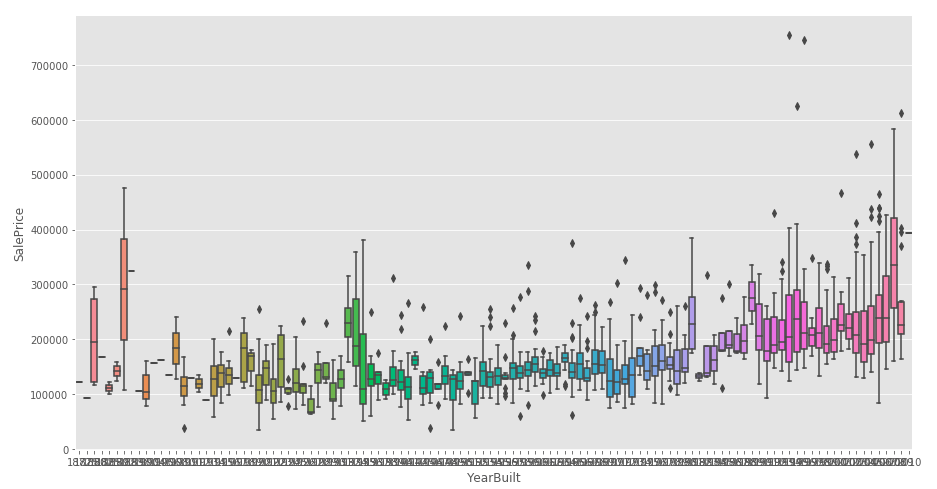
1. Python 3.7.4
2. Numpy 1.16.1
3. Pandas 0.25.1
4. Matplotlib 3.1.1
5. Scikit-learn 0.21.3
6. Spyder 4.1.3

## 模型——岭回归

岭回归(ridge regression)是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。

## 数据集——Kaggle房价预测数据集

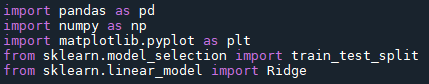
数据集使用79个特征描述一处住宅，根据这些特征对房价进行预测。以建造年份为例，进行可视化：



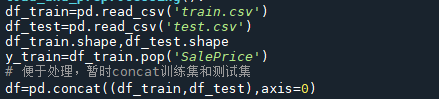
详见：<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview>

## 代码简介

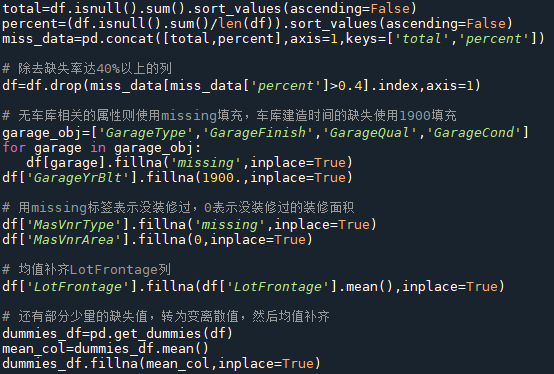
1. 导入包



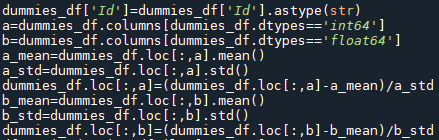
1. 读取数据集



1. 清洗数据



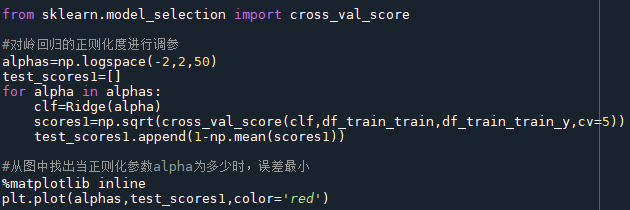
1. 标准化



1. 分割数据集



1. 岭回归调参



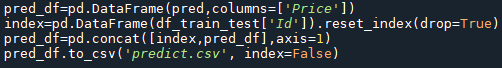
1. 训练模型



1. 计算效果



1. 保存预测值



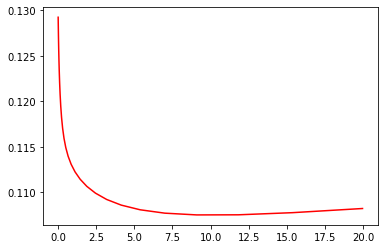
## 使用方法

1. 设置数据集路径



1. 对岭回归的正则化度进行调参

当岭回归的正则化参数alpha为10时，误差最小，因此后续训练将alpha设置为10



1. 训练

在Spyder中运行，进行训练，输出均方误差作为模型效果指标。



1. 保存预测结果predict.csv

