数据集

作业 1-3 将会使用同一个数据集，来自于 Kaggle 的房屋价格预测任务。

- ✧ 关于数据集每一列属性的描述可以参见 data_description.txt
- ✧ 我们的作业只使用 train.csv 文件里头的数据
- ✧ 数据集下载地址：
  https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

作业 2：分类与回归

作业提交内容：train_num.csv 和 validate_num.csv 文件，算**法代码和实验结果**

1. 【2 分】对 train.csv 的所有记录按照 8:2 的比例随机分成 training 和 validation 两个子数据集，并且每条记录只保留数值型属性，两个文件分别命名为 train_num.csv 和 validate_num.csv。

2. 【3 分】对 train_num.csv 文件训练一个 linear regression 模型对房屋价格进行预测，也就是说标签列是 SalePrice，其它属性当做输入向量。汇报在 validate_num.csv 的预测误差，包括 MAE 和 MSE。

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \qquad MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

3. 【5 分】将 train_num.csv 的 SalePrice 做离散化处理，每个数值对应的分类标签等于 x/100000 向上取整，例如房屋价格为 280000 的话，对应的 class label=3，然后对 train_num.csv 训练 SVM 和 Logistic Regression 模型，并汇报它们在 validate_num.csv 的正确率（即预测正确的数量除以总的预测次数），这里假设 validate_num.csv 文件中 SalePrice 标签也做同样的离散化处理。