# 项目1—房价预测问题

## Brief Introduction

你将得到所有房屋的共79个特征，这些特征包含了房屋的方方面面，比如建造年份，销售类型等。你的任务是根据这些特征使用线性回归模型对房屋的房价进行预测。这些特征有离散性特征也有连续性特征，同时也存在许多缺失的值(NA)，有些特征可能有用，有些特征可能无用，所以请认真对数据进行分析，构造合理的特征，再应用回归模型完成任务。

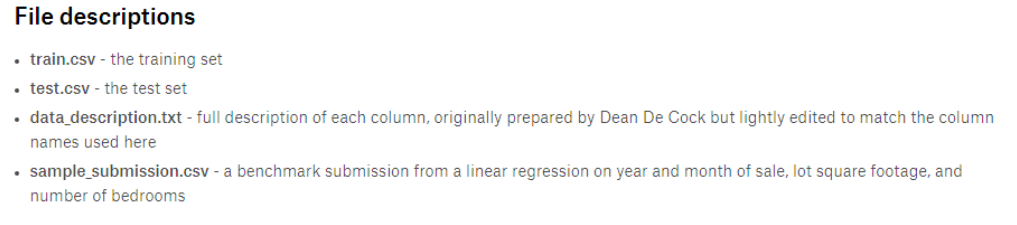
完成整个项目的步骤基本如下：

观察数据→数据清洗→重建特征→建立模型→调整参数→得到最终模型



## Dataset

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



* train.csv 训练集
* test.csv 测试集
* data\_description.txt 包含了对所有特征的解释

网页中提供了本题所需的数据集。csv是表格形式的文件，可以使用excel直接打开，每一列相当于一个特征（id列仅起到标识作用）。本练习包含两个数据文件，train.csv文件用于训练模型，test.csv文件用于评估模型。**注意：**在train.csv中，最后一列是SalePrice，代表我们要预测的房价值，这个就是我们的标签y，而在test.csv中是没有SalePrice这一列的，这个值需要使用我们训练好的模型去预测。

## Requirements

* **数据清洗和重建特征部分可以参考Kaggle上的一些做法，但是一定要明晰每一步处理的意义所在，不允许简单地复制其代码，避免学术不端的行为。**
* **本次项目限制必须使用线性回归的方法进行模型构建，搭建模型时不允许使用Sklearn等机器学习库中现成的线性回归函数，代价函数和梯度函数的代码细节需自己编写完成！**
* **项目采用分组形式：每组1~3人，项目里应包含代码，测试结果，各成员列表以及分工，发表所用的ppt等内容。**
* **项目材料最后按组各自打包提交。**

**Tips:** 自己手写的线性回归模型在优化的时候可能运行速度很慢，可以使用scipy提供的优化函数scipy.optimize.fmin\_tnc，具体用法参考[吴恩达逻辑回归](https://developer.aliyun.com/article/638717)(这是一个超链接)中的实现。这个函数可以理解为一个优化器，在你定义好自己的代价函数和梯度函数后可以使用这个优化器对参数空间进行优化。如果有疑问可以咨询助教。

## Evaluation

* 项目完成之后需要作**演示**，演示的内容包括特征处理部分，算法实现以及模型性能部分等。
* 鼓励使用可视化分析等手段对数据进行分析，也鼓励在线性回归模型的基础上添加更多预测手段以强化性能。**任何你觉得能让项目更好的功能，都可以作为加分项。**
* 代码需要成员**自主完成，严禁抄袭行为**，对自己团队代码的熟悉程度将一定程度影响最后评分。
* 由于本题限制了使用线性回归模型，所以性能上可能会低于Kaggle上的平均表现，但是这没有关系，只要性能不特别差都可以接受。本题更注重考察你的分析、建模能力，要有自己的思考和想法。