# 大数据期中 project

11510602 倪犀子
11510606 朱涵枫
11510618 孙磊

## 摘要

当下房价的飞速增长吸引了越来越多的人对房价的关注，我们知道一座房子的价格受到多方面因素的影响（地理位置，居住面积，房子质量等），如何结合多个因素对房子的价格进行客观评估，对于消费者在选购房子时有较大的帮助。本文根据已经搜集到的多方面会影响房价的数据，选定一部分作为训练样本，对这部分数据进行数据处理，可视化，最后通过构建回归模型，对测试集进行预测，得到这一部分数据的房价预测值，之后我们再将这些预测值与真实值进行比较，最后我们对结果进行评估，对模型进行评价。

## 一、 数据探索

1. **数据分析**：因为数据量较大,因此在进行具体的模型建立之前我们对数据进行了一些简单的数据分析，得到的具体结果见下表 1：

| 数据量 | 维数 | 自变量数 | 训练样本数 | 测试样本数 |
|---|---|---|---|---|
| 21614 | 21 | 19 | 15129 | 6485 |

表 1：数据的一些基本特征

（注：在计算时将 ID 去除，将 price 作为因变量）

2. **数据可视化**

①首先我们画出了因变量房价的分布图，具体图形见图 1：

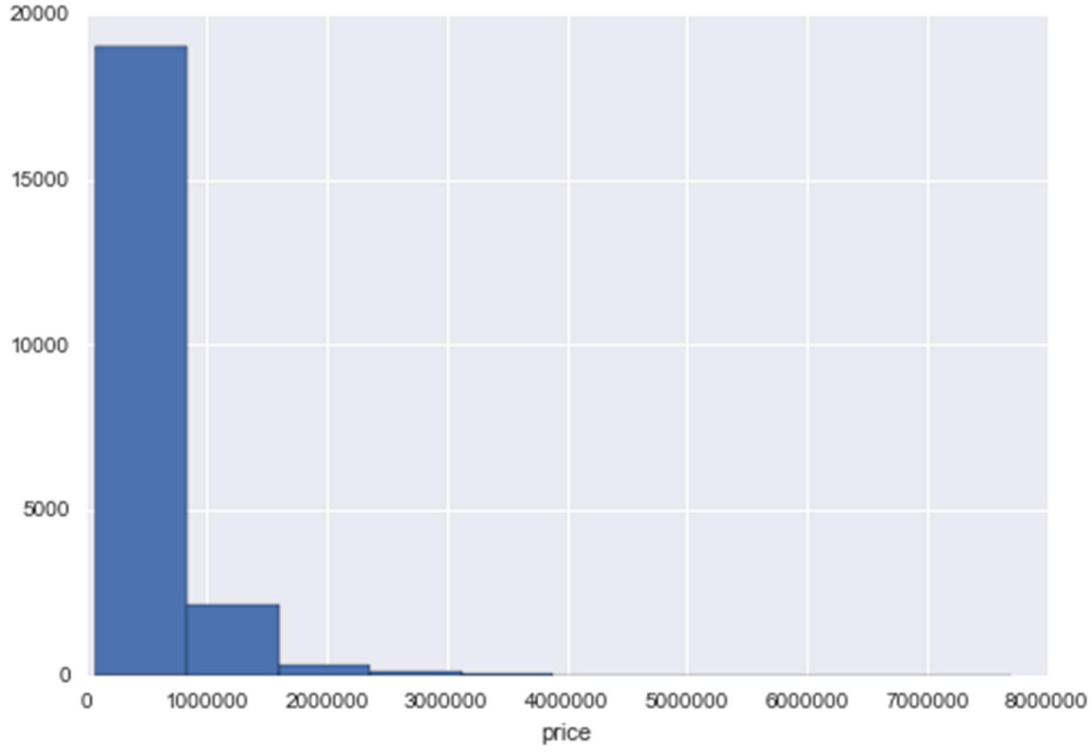

图 1：房价分布图

通过对图形的分析我们发现：数据的分布主要集中在 0-100w 之间，超过 200w 的数据之占很小一部分。

②之后我们做出了各自变量对于房价的变量箱体图，具体图形见图 2：（以 bedrooms、bathrooms 为例）

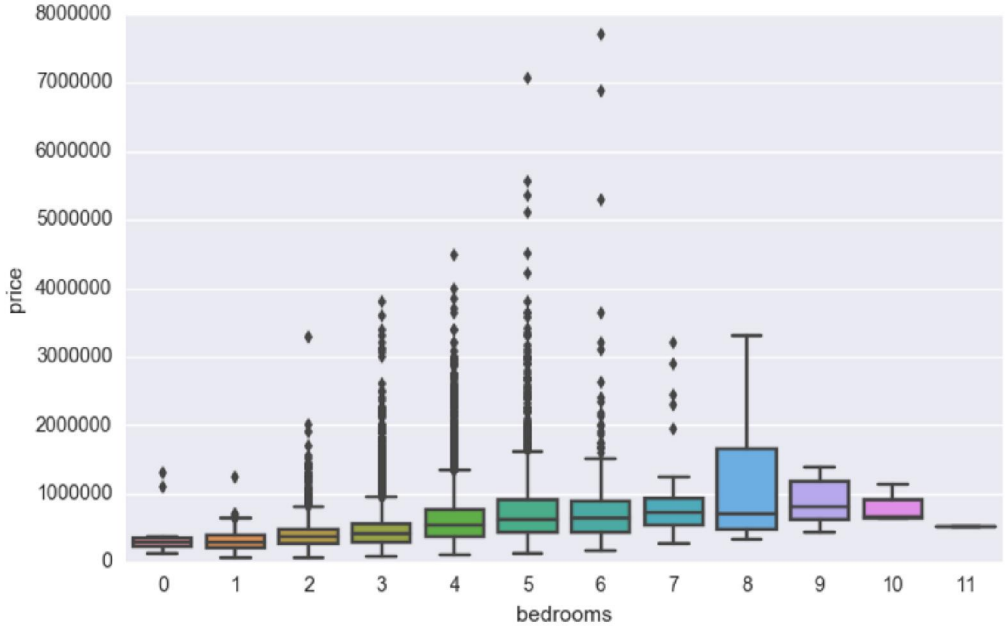


图 2：bedroom_price 箱体图

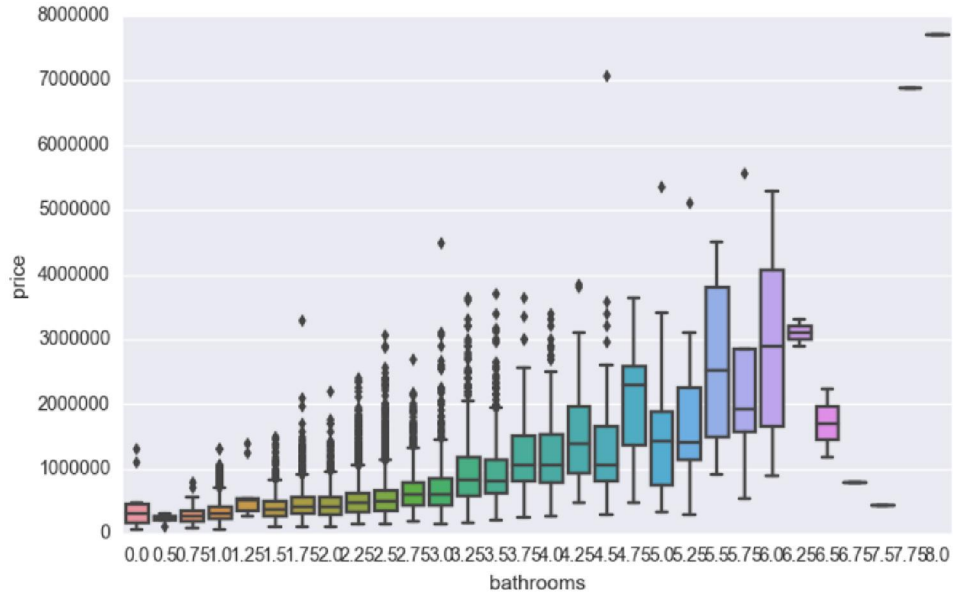


图 3：bathrooms_price 箱体图

箱体图的具体含义见下图 3：

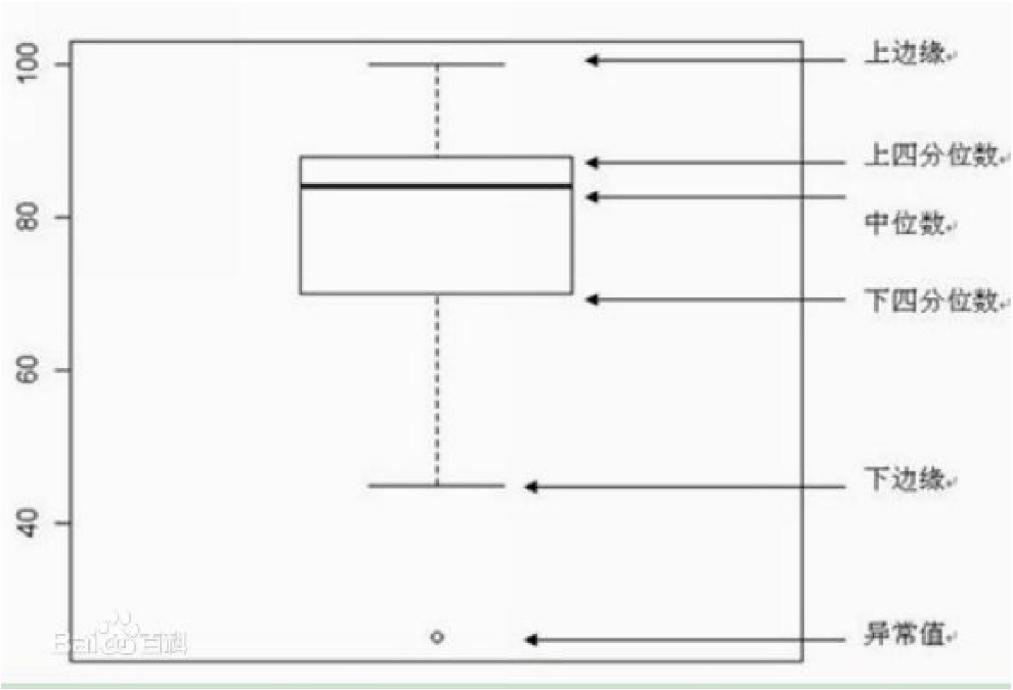


图 4：箱体图具体含义

散点图如下：（以 sqft_living、sqft_lot 为例）

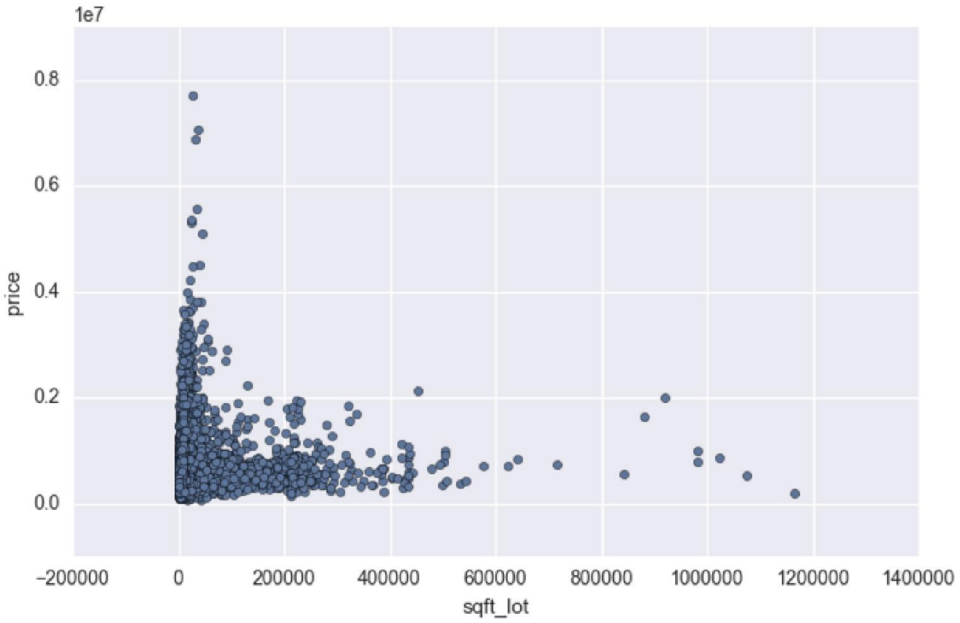


图 5：sqft_lot 散点图

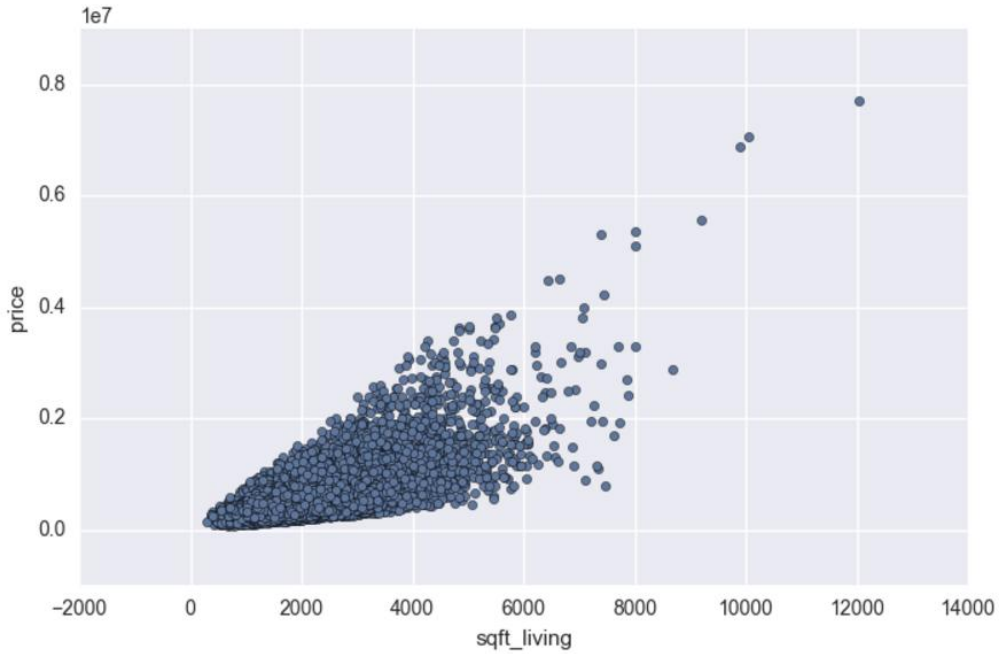


图 6：sqft_living 散点图

通过上图分析我们得到房价这一变量的数据分布如下表 2：

| 房价分布 | 数据 |
|---|---|
| 最大值（上边缘） | 7700000 |
| 最小值（下边缘） | 75000 |
| 上四分位数（75%） | 645000 |
| 下四分位数（25%） | 321950 |
| 中位数（50%） | 450000 |
| 均值（mean） | 5401822 |

表 2：房价数据分布

③随后做出变量相关性强度如下图 4：

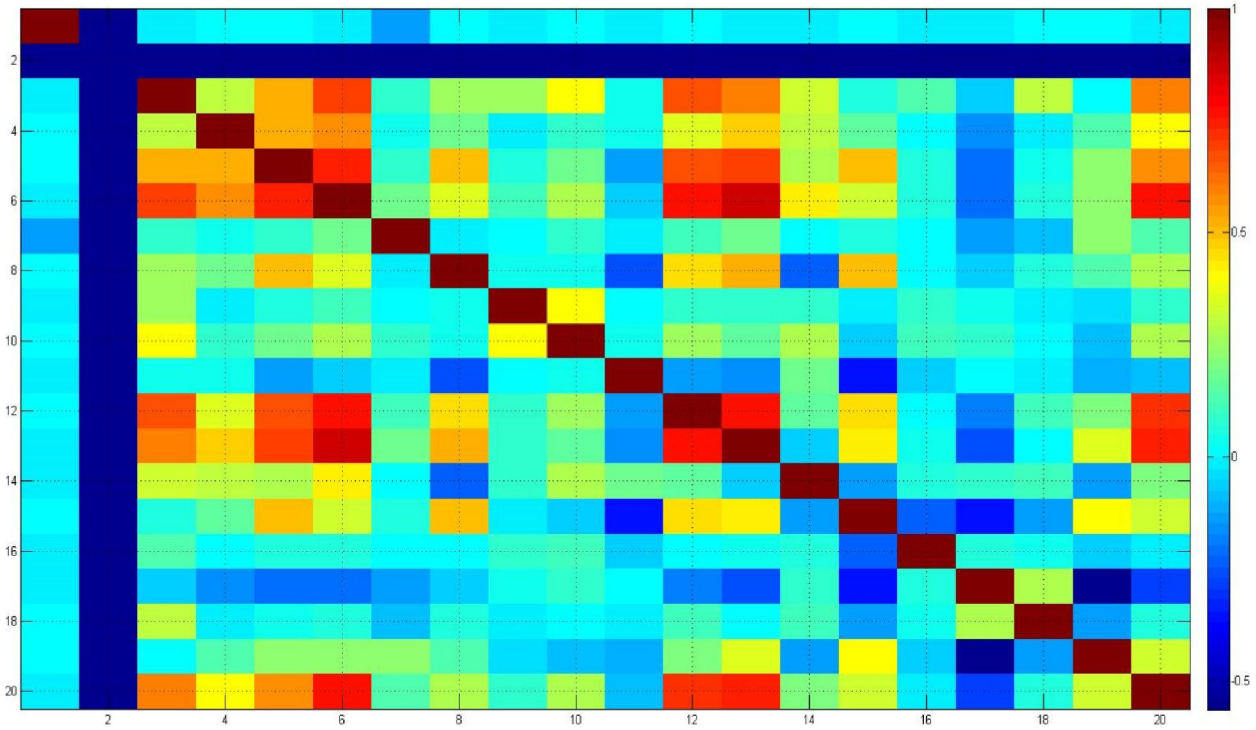

图 7：变量想关性强度图

④最后我们又绘制出了变量关联性如下图 5：

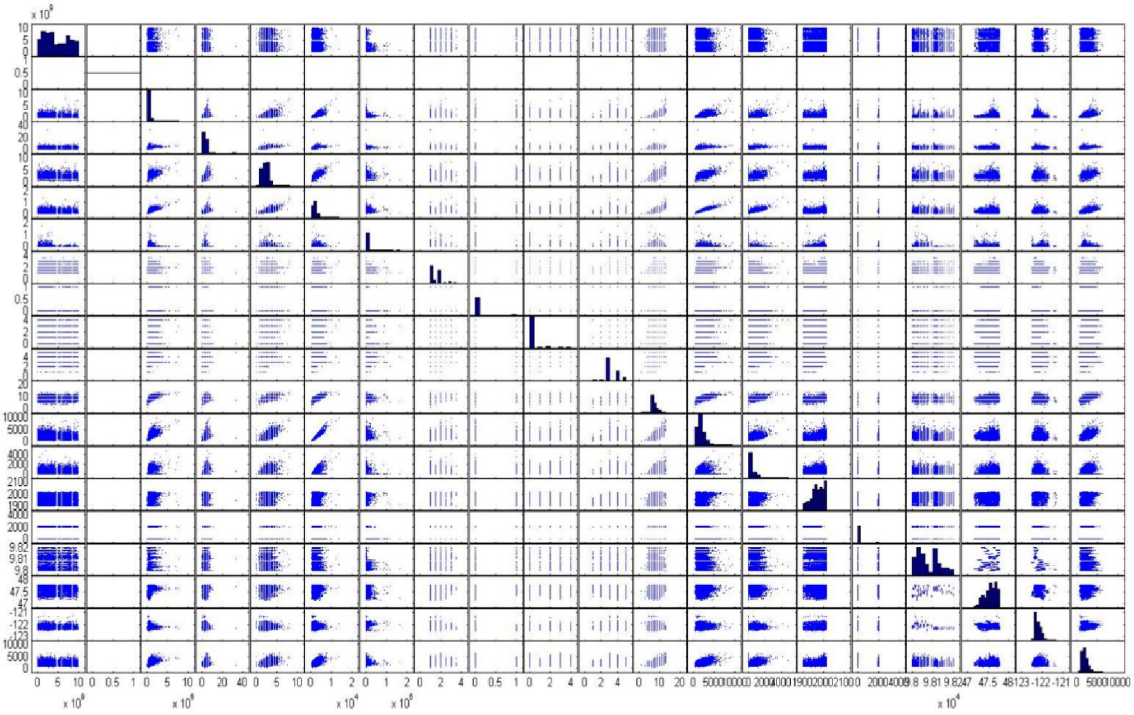

图 8：变量相关性关联图

## 二、 数据预处理

### 1. 训练集测试集划分

在具体操作中我们按照题目要求，将数据的前 70%（前 15129）个作为训练集，后 30%（后 6483）作为测试集

2. **缺失值处理**

通过 python 查找我们发现，数据中并没有缺失值。

3. **异常值处理**

首先我们将每一个变量所对应的数据分布打印了出来，之后对数据分布进行观察寻找离群点。最终得到异常值有如下 7 个：

| 编号 | 类别 | 数值 |
|---|---|---|
| 1 | bedrooms | 33 |
| 2 | sqft_lot | 1651359 |
| 3 | sqft_basement | 4130 |
| 4 | sqft_basement | 4820 |
| 5 | sqft_lot5 | 560617 |
| 6 | sqft_lot5 | 858132 |
| 7 | sqft_lot5 | 871200 |

表 3：数据异常值

4. **数据标准化**

在标准化时，我们采用的是 0-1 标准化，在具体过程中我们去掉了 date 和 price 这两个变量，得到其余 17 个变量最大值，最小值如下表 3：

| 0-1 标准化 | | |
|---|---|---|
| 变量名 | 最大值 | 最小值 |
| bedrooms | 1 | 0 |
| bathrooms | 1 | 0 |
| sqft_living | 1 | 0 |
| sqft_lot | 1 | 0 |
| floors | 1 | 0 |
| waterfront | 1 | 0 |
| view | 1 | 0 |
| condition | 1 | 0 |
| grade | 1 | 0 |
| sqft_above | 1 | 0 |
| sqft_basement | 1 | 0 |
| yr_built | 1 | 0 |
| yr_renovated | 1 | 0 |
| zipcode | 1 | 0 |
| lat | 1 | 0 |
| long | 1 | 0 |
| sqft_living15 | 1 | 0 |
| sqft_lot15 | 1 | 0 |

表 4：0-1 标准化后各变量统计表

标准化后我们对部分数据的具体截图如下：

| Index | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 0.0909 | 0.125 | 0.0672 | 0.00311 | 0 |
| 1 | 0.0909 | 0.281 | 0.172 | 0.00407 | 0.4 |
| 2 | 0.0606 | 0.125 | 0.0362 | 0.00574 | 0 |
| 3 | 0.121 | 0.375 | 0.126 | 0.00271 | 0 |
| 4 | 0.0909 | 0.25 | 0.105 | 0.00458 | 0 |
| 5 | 0.121 | 0.562 | 0.387 | 0.0614 | 0 |
| 6 | 0.0909 | 0.281 | 0.108 | 0.00382 | 0.4 |
| 7 | 0.0909 | 0.188 | 0.0581 | 0.00557 | 0 |
| 8 | 0.0909 | 0.125 | 0.112 | 0.00421 | 0 |
| 9 | 0.0909 | 0.312 | 0.121 | 0.00366 | 0.4 |
| 10 | 0.0909 | 0.312 | 0.247 | 0.00562 | 0 |
| 11 | 0.0606 | 0.125 | 0.0657 | 0.00332 | 0 |
| 12 | 0.0909 | 0.125 | 0.086 | 0.0117 | 0.2 |
| 13 | 0.0909 | 0.219 | 0.0815 | 0.00555 | 0 |
| 14 | 0.152 | 0.25 | 0.115 | 0.00262 | 0.2 |
| 15 | 0.121 | 0.375 | 0.201 | 0.00271 | 0.4 |
| 16 | 0.0909 | 0.25 | 0.121 | 0.00819 | 0.4 |

图 9：标准化后具体数据截图

# 三、 变量处理选择与模型构建

### ①变量的分析与处理

首先我们对日期 date 这一字符串变量进行了数值化处理，在具体的处理过程中我们选择采用 2014.5.1 作为第一天，依此类推对日期进行排序编号。

之后我们对变量进行分析，通过之前做出的散点图我们发现部分变量与房价有着明显的线性关系，但是有一些与房价更符合其他函数关系，我们对这些与房价有着其他函数关系的变量进行了汇总，通过不断测试我们找到了能使预测结果最好的具体函数关系，结果如下表所示：

| 变量名 | 与房价函数关系 |
|---|---|
| **Grade** | 4.7 次方 |
| **Bathroom** | 1.000000001 次方 |
| **Lat** | 8.6 次方 |
| **Sqft_lot** | -0.3 次方 |
| **Yr_built** | 4.1 次方 |

表 5：部分变量与房价函数关系表

**②线性回归、Lasso 回归、岭回归的构建**

线性回归主要是通过极小化残差平方和的方法来求得系数节解点从而拟合出线性方程。

Lasso 和岭回归在线性回归的基础上加一个 L2 或 L1 正则化项，来控制模型空间，减小模型复杂度，防止过拟合。

可以看出他们对于变量的要求都是尽量满足线性关系，因此我们在构建时主要还是围绕变量展开，过程如下：

我们对这些与房价不符合线性关系的变量，进行了函数处理，然后用这些变量来对房价进行预测。在划分训练集时，我们将前 70%的数据作为训练集，后 30%的数据作为测试集，我们选择采用 $R^2$，mae，mse，rmse 这四个评价指标来对我们的模型进行评价。

**③随机森林的构建**

随机森林：随机的进行属性和样本的选取，生成多个各不相同的决策树对样本进行预测，最后汇总决策树的结果得出最终的结论。

由于随机森林并不需要线性化，并且原始数据维数较小不适合进行降维处理。因此我们没有做对变量做太多处理，只是将 id 这一变量去除，将 date 这一变量进行了数值化处理。

**④BP 人工神经网络的构建**

BP 人工神经网络是由大量的节点（或称神经元）和之间的相互联接构成，每个节点代表一个特定的输出函数（称为激励函数），每两个节点之间的联接都代表一个对于通过该联接信号的加权值（称为权重），权重相当于人工神经网络的记忆。网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。

（注：通过测试我们发现利用标准化后的数据效果并不是特别好，因此在具体运行中我们利用的还是原始数据）

# 四、 模型测试

**1.线性回归、Lasso 和岭回归**

对数据预处理后所得数据划分训练集，测试集，用 python 进行拟合回归，用测试集的样本特征来进行预测，与测试集的实际值比较，求出 $R^2$ 等具体评价数据基本都在 0.7 左右，并且调整模型参数对结果影响不大（以上内容在之后几个模型中就不复述了）。

**2.随机森林**

由于随机森林在运行时占用内存较大，因此我们在具体测试时，进行了一下参数调整将最大深度（max_depth）减小，这样使得我们可以较大程度的增加树的个数，也能得出较为不错的结果。

**3.BP 人工神经网络**

我们运用 matlab 构建 BP 人工神经网络，网络分为输入层--隐藏层--输

出层，依次的节点个数取 18-13-1。训练次数为 600 次。
具体计算评价指标的步骤为：
① 利用 Matlab 得出对于房价的预测值（储存在 anew 这一数组中）
② 之后我们将这些预测值导入到了一个 excel 文件中（具体文件名为：kc_house_data3.csv）
③ 最后我们利用 python 读取这些数据，计算得到这些评价指标的具体数值。

**4.具体运行结果：**

| 评价指标 | 线性回归 | 岭回归 | Lasso 回归 | 随机森林 | BP 人工神经网络 |
|---|---|---|---|---|---|
| R2_score | 0.7461173491 | 0.7356922422 | 0.7361493393 | 0.73208959 | 0.838120735 |
| Mean absolute error | 116151.042849 | 119466.13547 | 119042.68302 | 102730.208 | 89745.27093 |
| Mean squared error | 34433741378.3 | 35847683740 | 35785688307 | 36336306192 | 20691361755 |
| Root mean squared error | 185563.308276 | 189334.845551 | 189171.05557 | 190620.84406 | 143844.9226 |

**表 6：各模型具体运行结果**

# 五、 模型评价

1.线性回归、Lasso 和岭回归
由三者相差无几的结果可知，训练模型不存在过拟合的问题，虽然我们对一些变量进行了线性化处理，但是仍然有部分变量(如日期、经纬度等)的部分样本特征与房价无明显线性关系，因此仅仅使用用线性方程去拟合预测房价是不会得到较好的结果的，而且从不太好的 R2 值也可看出，房价的预测不太适合用这三个模型解决。
2.随机森林
随机森林在预测结果上不是特别好，除此之外它的解释性也不高，我们对此具体的分析如下：
分析数据我们发现对于房价的各个属性划分较细（如：sqft_living15 等），取值划分较多，这些属性在一定程度上会对随机森林的预测产生较大影响，使得预测结果不会太好。
3.BP 人工神经网络
BP 神经网络是模仿人脑学习过程的一种模型，其中的激励函数、权重与随机森林中决策树分支结点（熵、信息增益率、Gini 不纯度等）有着一些类似的性质，要么都是在模型训练中确定数值，要么都是决定输出结果的重要参数或者函数，最关键的是，都不能给出房价与各个样本特征之间具体的联系。不难明白，BP 人工神经网络的解释性也较差。但是，BP 人工神经网络模型对内存要求不高，在解决房价预测的问题上也有很好的效果。我们通过这个模型，得出了我们所尝试的所有模型的结果中最好的一个结果。

# 六、 小组具体分工

**朱涵枫：** 数据分析、数据可视化、数据预处理、论文书写

**孙磊：**随机森林模型构建与测试、神经网络模型构建与测试、变量特征处理

**倪犀子：** 线性回归、Lasso 回归、岭回归模型构建与测试、模型测试