网格化存量房评估模型简介

# 研究过程

## 数据采集

### 片区边界数据

#### 行政区划

行政区划数据来自电子地图。

#### 土地级别

土地级别数据来自国土局公布的官方数据。

#### 街道

街道数据使用专业的GIS软件QGIS进行绘制。

#### 学区

学区数据来自安徽各市教育局公布的官方数据。

### 影响因素数据

影响因素主要包括

## 数据分析

### 相关性分析

### 等级划分

对于有明确等级的因素，如土地级别、医院、景区等，直接以其对应的等级作为因素等级。而对于没有明确等级的因素，如学区、商场等，结合实地调查得到的规模、口碑等因素进行人为分级。

### 主成分分析

主成分分析是一种常见的数据降维手段，顾名思义，主成分分析就是找到数据中的最主要的方面来代替原始数据，能够起到降低数据噪声、去除冗余数据的作用。具体操作步骤如下：

#### 对样本数据进行中心化，即将样本中每个因素的均值置为0，；

#### 计算样本数据的协方差；

#### 求出协方差矩阵的特征值即对应的特征向量；

#### 将特征向量按照特征值大小进行排列成矩阵，取前k行组成矩阵P；

#### 即为将原始数据降到k维之后的结果。

## 数据处理

### 缺失值处理

在数据采集过程中难免会出现缺失值，如部分地区因为学校在建或者学区规划尚未公布而没有学区，这样就导致学区数据为NAN(not a number)。

### 偏态数据处理

线性回归理论中，模型的预测值和实际值之间会存在一个残差项： ，当这个残差项符合正态分布时才能获得一个无偏的模型。与之对应的训练数据X，预测目标Y都应该符合正态分布才能使模型更加准确。因此，在建立线性回归模型之前，对偏态分布的数据进行处理是很重要的一个步骤。

### 离群点处理

线性模型追求的是总体的误差最小，而原始数据中常常会出现离群点，比如某个市中的某个小区可能会出现一个房价远超该市均价的小区，为了使总体的误差最小，线性模型会偏向这种异常小区数据，所以这种小区会导致模型中对与该小区相似的小区价格被显著高估，导致模型失准，所以在训练模型时，要先去除这一类数据。

# 模型简介

## 岭回归模型

## 随机森林模型

# 模型效果

## 小区模型

## 学区模型