大数据智能分析理论与方法

课程设计报告

班级：

学号：

姓名：

# 线性回归分析租房价格

**摘要：**

本项目旨在通过分析租房数据，运用线性回归模型预测租房价格。通过数据预处理、模型构建、模型分析和可视化展示，为租房市场提供价格预测和决策支持。

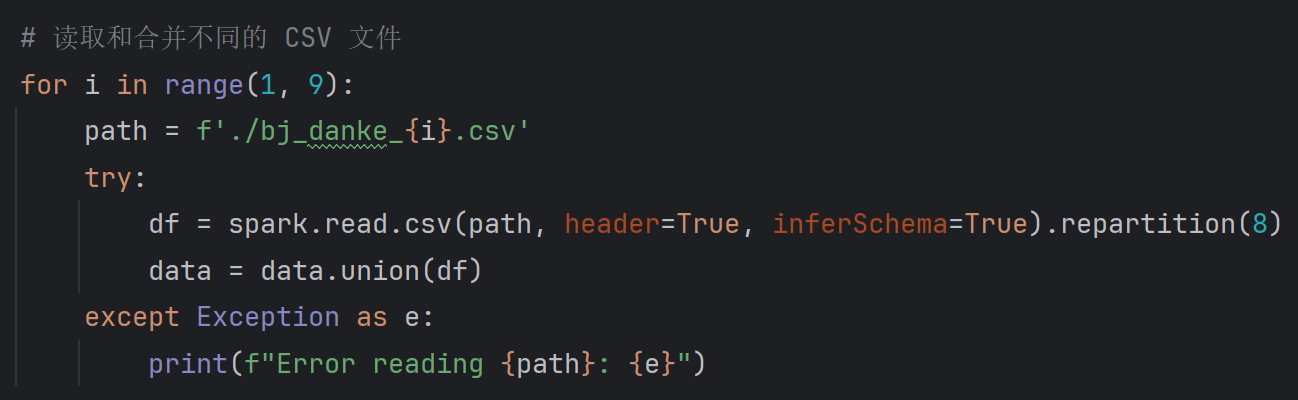
**引言：**

本项目通过PySpark平台成功构建了线性回归模型，对租房价格进行了有效预测。模型的准确性和稳定性得到了验证，为租房市场提供了有价值的参考。未来可以通过收集更多数据和尝试不同的算法来进一步提高模型的预测能力。

## 第一章 数据收集与预处理

### 1.1 数据收集

本项目收集了北京市租房市场的公开数据，包括房屋的地理位置、类型、面积、装修情况等信息。数据来源于多个CSV文件，分别为bj\_danke\_1.csv至bj\_danke\_8.csv。



### 1.2 数据预处理

数据预处理是数据分析中的重要步骤，其目的是确保数据的质量和一致性。本项目中，我进行了以下预处理操作：

* **去重**：删除数据中的重复记录，确保每条记录的唯一性。
* **数据清洗**：删除包含任何缺失值的行，以避免在后续分析中引入错误。
* **数据转换**：对数据进行必要的转换，例如将文本类型的数据转换为整数类型。

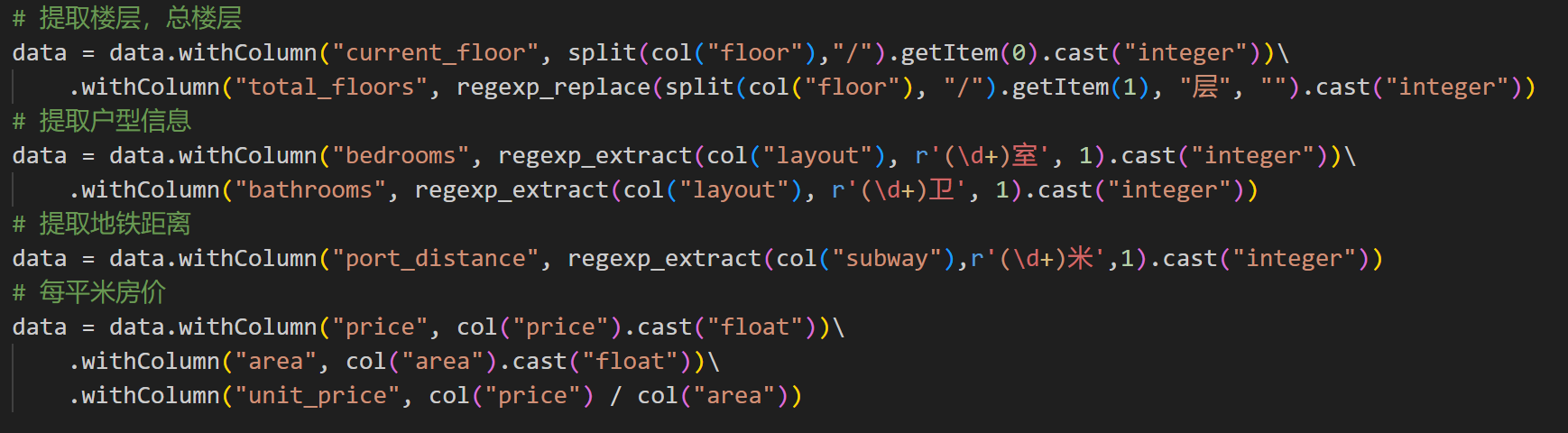
文本

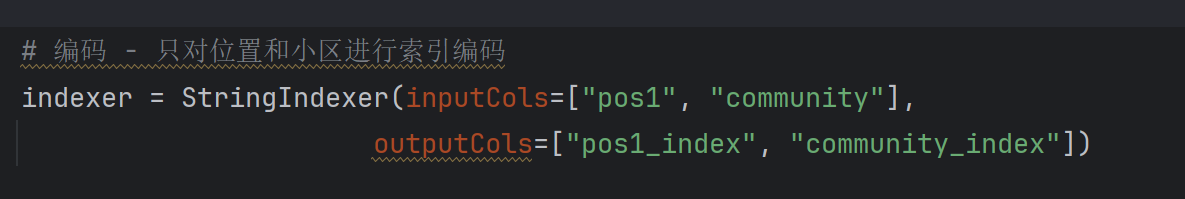
描述已自动生成

### 1.3 数据提取

在数据预处理阶段，我对原始数据进行了转换和提取操作：

* **提取楼层信息**：从floor字段中提取出当前楼层(current\_floor)和总楼层(total\_floors)。
* **提取户型信息**：从layout字段中提取出卧室数量(bedrooms)和卫生间数量(bathrooms)。
* **提取地铁距离**：从subway字段中提取出地铁距离(port\_distance)。
* **计算每平米房价**：计算并添加了单位面积价格(unit\_price)字段。





## 第二章 模型构建与分析

### 2.1 模型构建

我使用了PySpark的MLlib库构建了线性回归模型，以预测租房价格。模型的构建包括以下步骤：

* **特征工程**：使用VectorAssembler将多个特征组合成一个特征向量，并使用StringIndexer和OneHotEncoder对类别特征进行编码。
* **模型训练**：使用线性回归算法训练模型，并使用CrossValidator进行参数调优。

文本

描述已自动生成

文本

描述已自动生成

### 2.2 模型分析

模型分析的目的是评估模型的性能，并进行必要的调整和优化。本项目中，我进行了以下分析：

* **模型评估**：使用均方根误差(RMSE)评估模型的预测性能。
* **残差分析**：分析残差的分布，以检查模型是否满足线性回归的假设。

文本

描述已自动生成

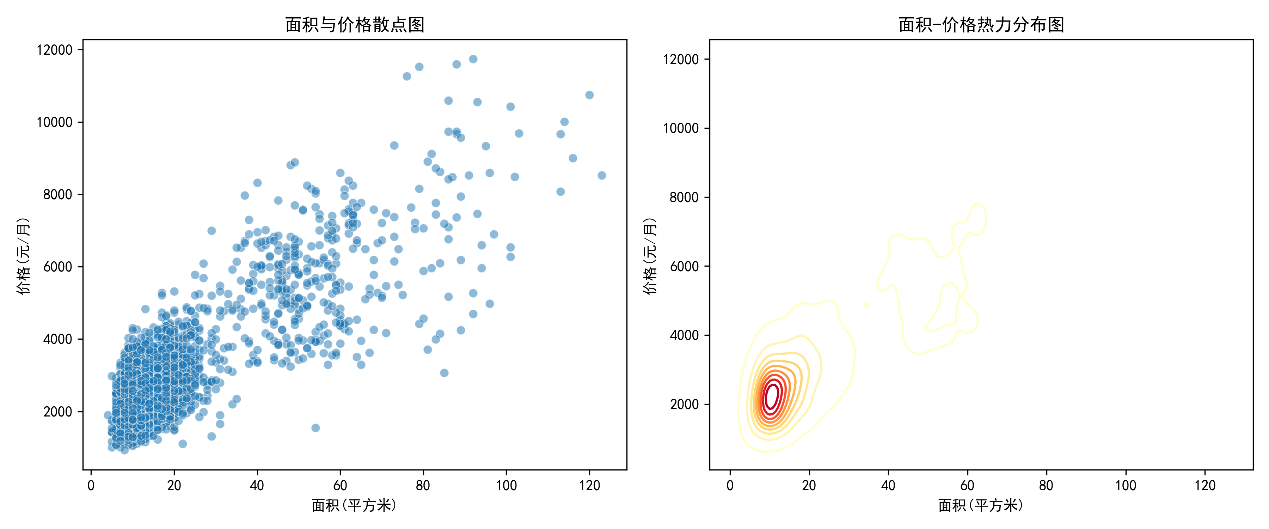
文本

描述已自动生成

## 第三章 可视化展示

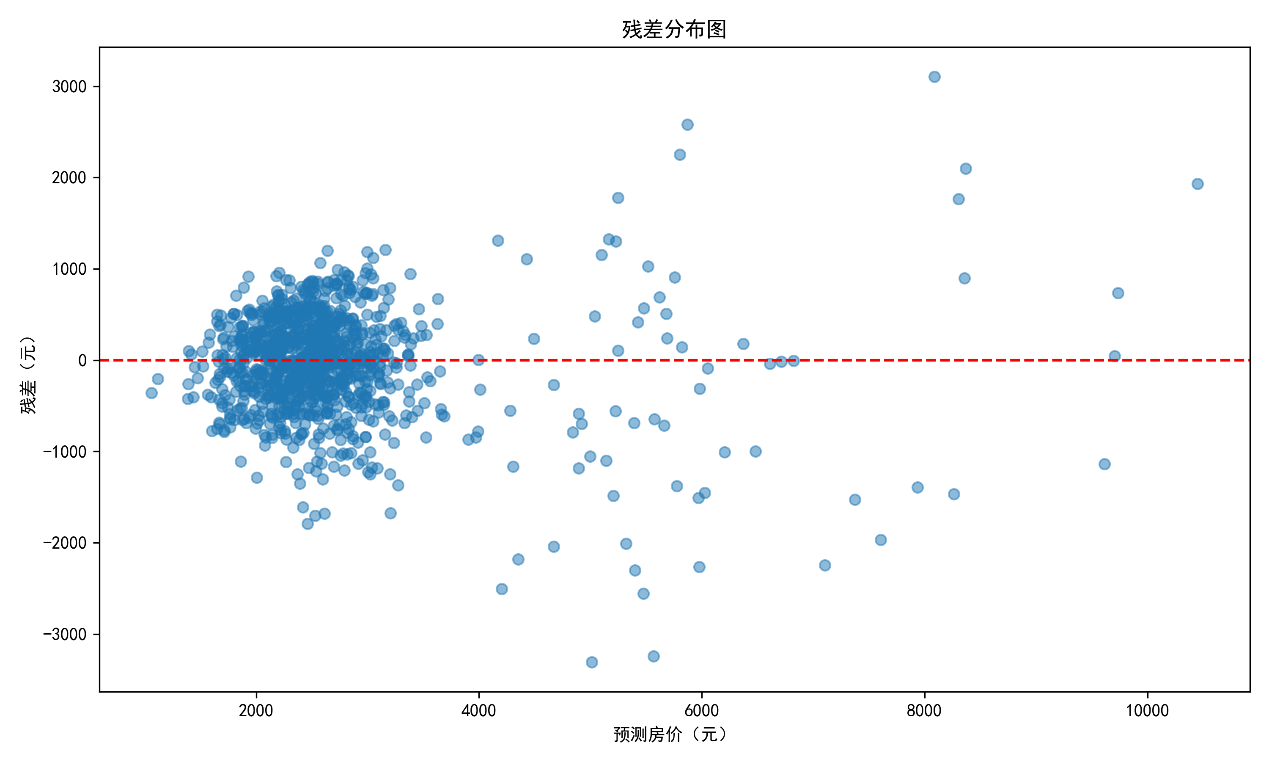
### 3.1 价格分布图

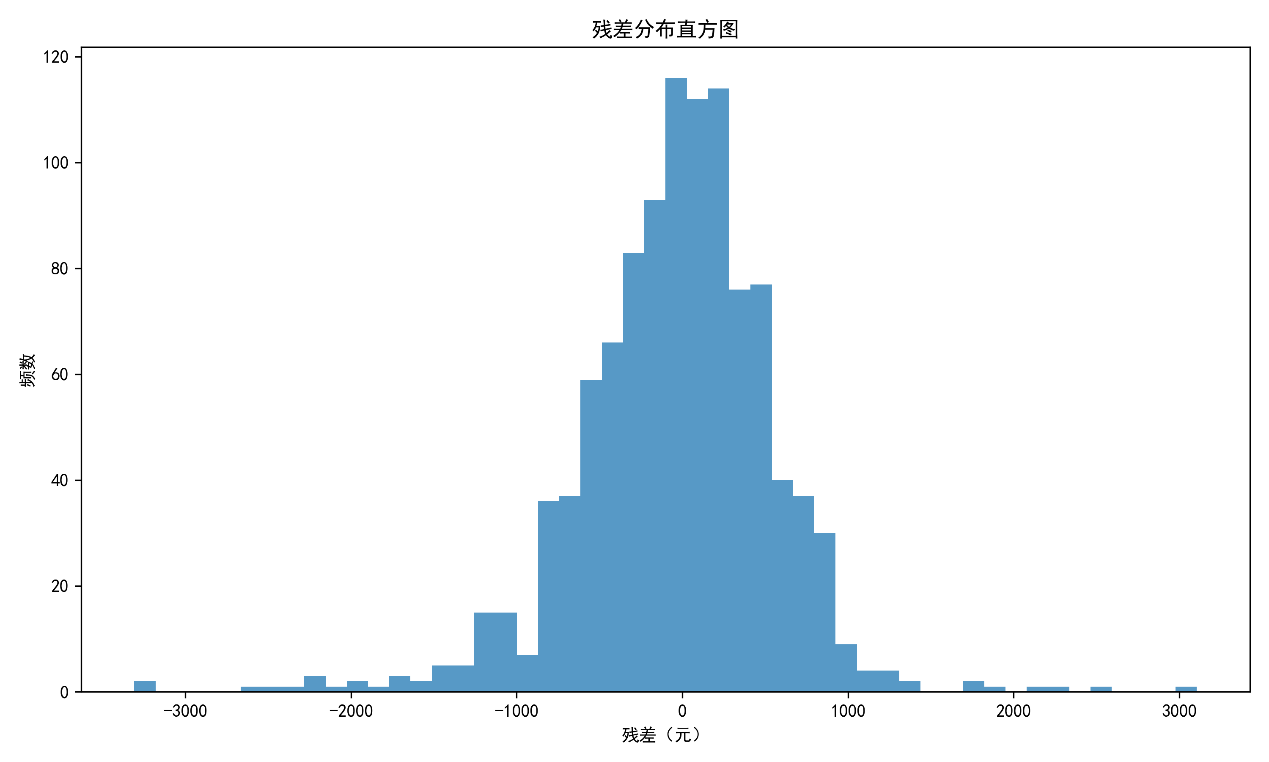
使用散点图展示了模型预测的租房价格与实际价格的对比，并通过添加对角线来直观展示预测值与实际值的一致性。



### 3.2 残差分布图

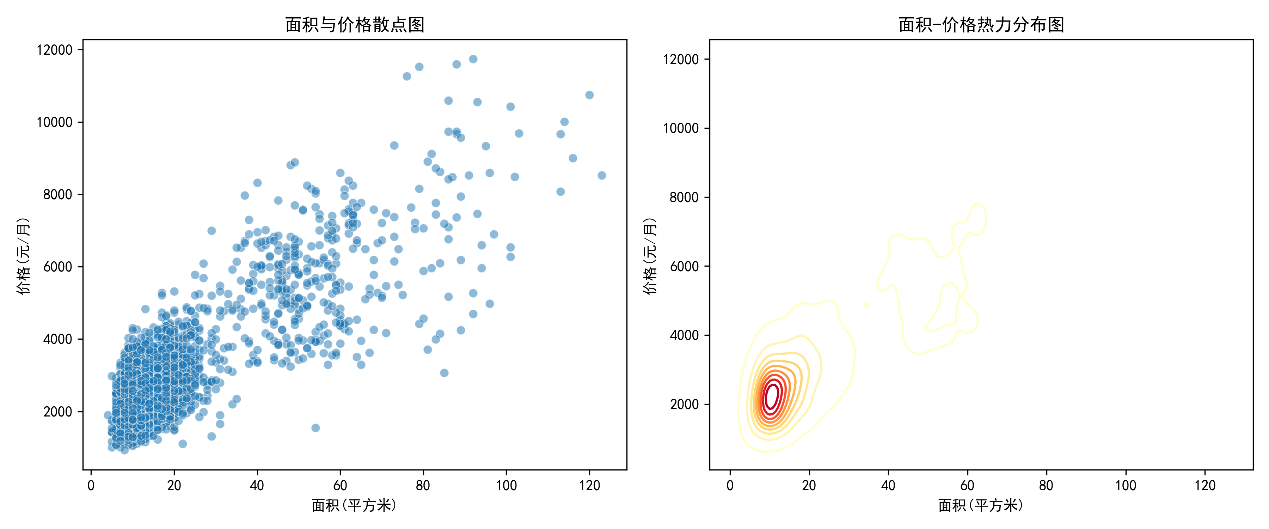
绘制了残差的散点图和直方图，以检查残差的分布是否满足正态分布的假设。





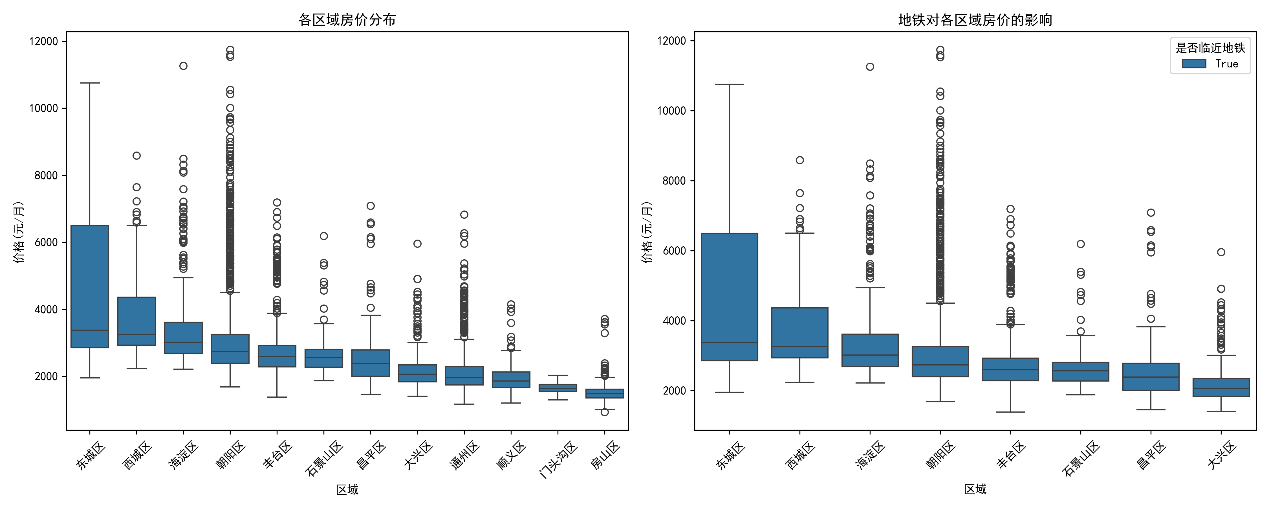
### 3.3 面积与价格关系分析图

使用散点图和热力图展示了面积与价格的关系，以及不同面积下价格的分布情况。



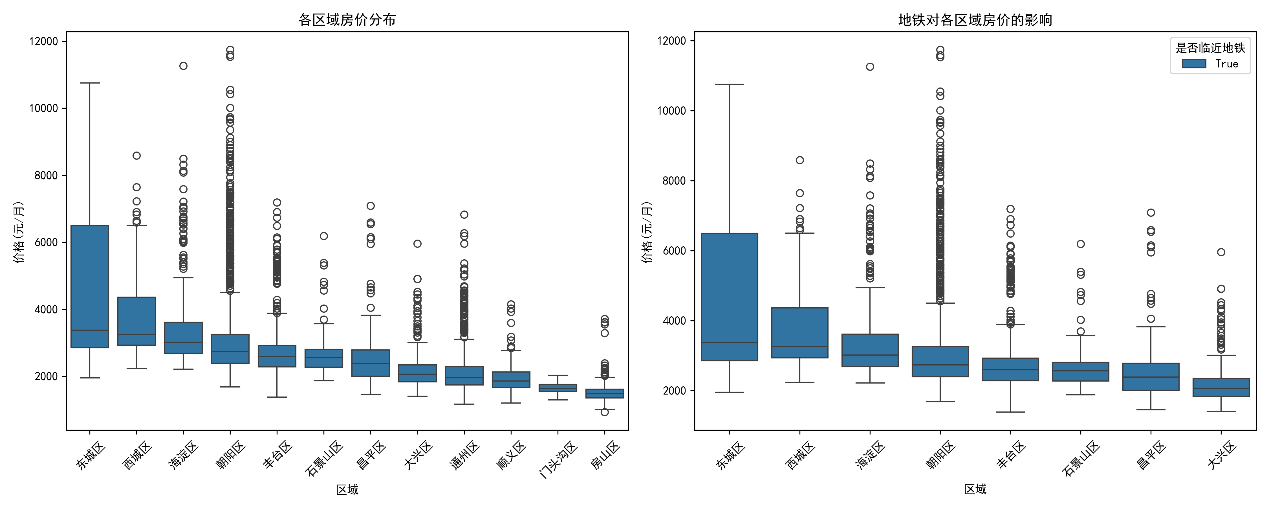
### 3.4 位置与地铁分析图

使用箱线图展示了不同区域的房价分布，以及地铁对房价的影响。



### 3.5 户型分析图

使用小提琴图和箱线图展示了不同户型的房价分布，以及单位面积价格的分布情况。



### 3.6 小区分析图

使用条形图展示了热门小区的平均房价对比，并标注了各小区的房屋数量。