

**实 验 报 告**

|  |  |  |
| --- | --- | --- |
| **课程名称** | **:** | **人工智能程序设计** |
| **实验名称** | **:** | **实验1 房价预测：数据预处理与可视化** |
| **专业班级** | **:** | **2022级计算机科学与技术1班** |
| **姓名** | **:** |  |
| **学号** | **:** |  |
| **指导老师** | **:** | **林耀威** |
| **日期** | **:** | **2024年11月21日** |

**福州大学至诚学院**

实验1 房价预测：数据预处理与可视化

### 实验目的与要求

1. 加深对数据认识、数据探索、数据预处理的理解和认识；
2. 掌握DataFrame对数据进行探索、预处理方法；
3. 掌握DataFrame、Matplotlib的可视化探索。

### 实验准备

1. 环境：Python3、Pycharm 社区版或Jupyter Notebook；
2. 复习DataFrame对数据的探索分析的API方法，诸如info、replace、 fillna、map、apply等；
3. 学习DataFrame、Matplotlib对数据的可视化。

### **实验数据集**

亚利桑那州埃姆斯市的房价数据集是由Bart de Cock于2011年收集， 涵盖了2006-2010年期间亚利桑那州埃姆斯市的房价，这个数据集比哈里森和鲁宾菲尔德的波士顿房价数据集要大得多，也有更多的特征。每条记录包括了房屋的属性，如街道类型、施⼯年份、屋顶类型、地下室状况等。这些特征由各种数据类型组成。 例如，建筑年份由整数表示，屋顶类型由离散类别表示，其他特征由浮点数表示。 kaggle比赛链接如下:

[https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview" \t "http://localhost:8889/notebooks/_blank)

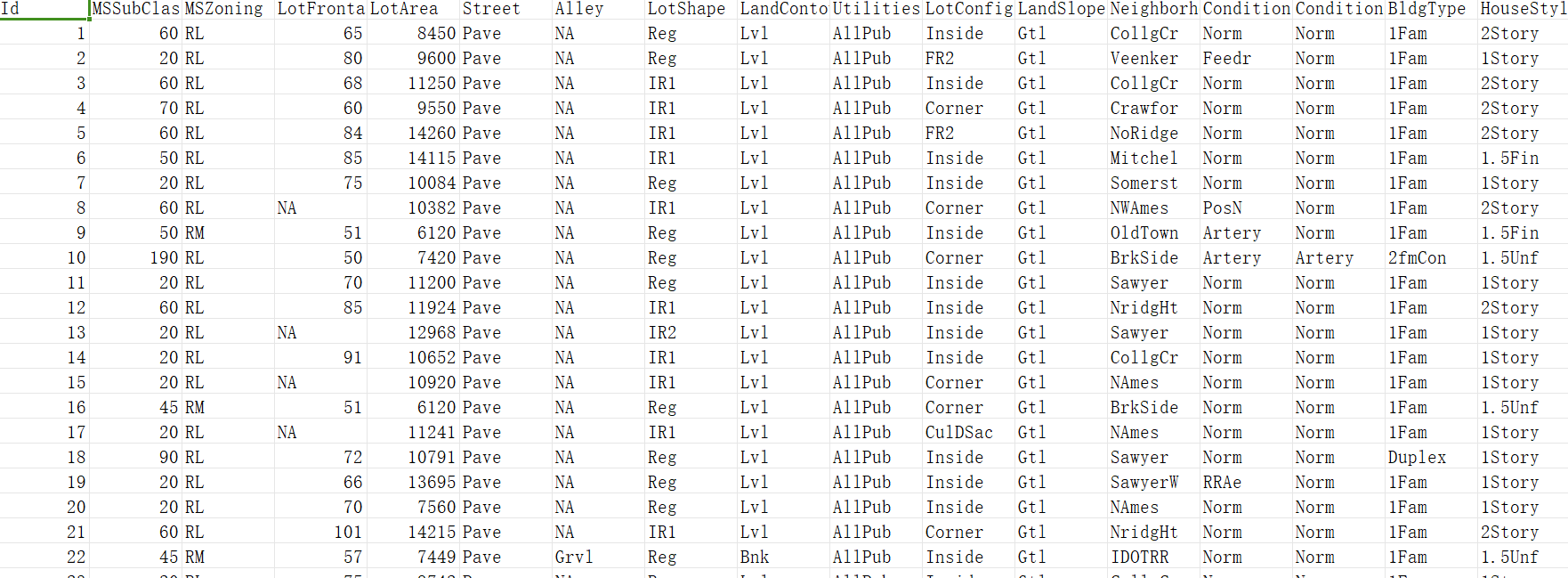


图1 房价数据集训练集部分列示例

训练数据集包括1460个样本，每个样本80个特征和1个标签， 而测试数据集包含1459个样本，每个样本80个特征。80个特征和标签的说明如下表所示：

|  |  |
| --- | --- |
| 属性字段 | 描述 |
| Id | 交易ID，唯一编码 |
| MSSubClass | 建筑类型 |
| MSZoning | 分区类型 |
| LotFrontage | 与房相连的街道长度（英尺） |
| LotArea | 地表面积（平方英尺） |
| Street | 道路通行类型 |
| Alley | 胡同通道类型 |
| LotShape | 地基形状 |
| LandContour | 房子平整度 |
| Utilities | 公用面积类型 |
| LotConfig | 地基配置 |
| LandSlope | 房子坡度 |
| Neighborhood | 邻居 |
| Condition1 | 靠近主要道路或铁路 |
| Condition2 | 靠近主要道路或铁路（标准2） |
| BldgType | 住宅类型 |
| HouseStyle | 住宅风格 |
| OverallQual | 整体材料和表面质量 |
| OverallCond | 总体状况评级 |
| YearBuilt | 原建造日期 |
| YearRemodAdd | 改造日期 |
| RoofStyle | 屋顶类型 |
| RoofMatl | 屋顶材料 |
| Exterior1st | 房子的外部覆盖材料 |
| Exterior2nd | 房屋的外部覆盖物（如果超过一种材料） |
| MasVnrType | 砖石饰面类型 |
| MasVnrArea | 砖石贴面面积（平方英尺） |
| ExterQual | 外观材质 |
| ExterCond | 外部材料现状 |
| Foundation | 地基类型 |
| BsmtQual | 地下室高度 |
| BsmtCond | 地下室状况 |
| BsmtExposure | 花园地下室墙壁 |
| BsmtFinType1 | 地下室竣工面积质量 |
| BsmtFinSF1 | 完成类型1（平方英尺） |
| BsmtFinType2 | 第二个成品区域的质量（如果存在） |
| BsmtFinSF2 | 完成类型2（平方英尺） |
| BsmtUnfSF | 未完工的平方英尺面积 |
| TotalBsmtSF | 地下室面积总平方英尺 |
| Heating | 供热类型 |
| HeatingQC | 供热质量及条件 |
| CentralAir | 中央空调条件 |
| Electrical | 电气系统 |
| 1stFlrSF | 一楼面积（平方英尺） |
| 2ndFlrSF | 二楼面积（平方英尺） |
| LowQualFinSF | 低质量成品平方英尺（所有楼层） |
| GrLivArea | 地上（地面）居住面积平方英尺 |
| BsmtFullBath | 地下室设施齐全浴室 |
| BsmtHalfBath | 地下室设施半浴室 |
| FullBath | 地上室设施齐全浴室 |
| HalfBath | 地上室设施半浴室 |
| BedroomAbvGr | 卧室数量 |
| KitchenAbvGr | 厨房数量 |
| KitchenQual | 厨房面积 |
| TotRmsAbvGrd | 房间总数（不包括浴室） |
| Functional | 家居功能评级 |
| Fireplaces | 壁炉数量 |
| FireplaceQu | 壁炉品质 |
| GarageType | 车库位置 |
| GarageYrBlt | 车库建成年份 |
| GarageFinish | 车库的内部装修 |
| GarageCars | 车库大小（汽车容量） |
| GarageArea | 车库大小（平方英尺） |
| GarageQual | 车库质量 |
| GarageCond | 车库状况 |
| PavedDrive | 铺好的车道 |
| WoodDeckSF | 木板面积（平方英尺） |
| OpenPorchSF | 开放式门廊面积（平方英尺） |
| EnclosedPorch | 封闭式门廊面积（平方英尺） |
| 3SsnPorch | 三季阳台（平方英尺） |
| ScreenPorch | 屏风门廊面积（平方英尺） |
| PoolArea | 泳池面积（平方英尺） |
| PoolQC | 泳池质量 |
| Fence | 护栏网质量 |
| MiscFeature | 其他类别未涵盖的杂项功能 |
| MiscVal | 杂项特征的值 |
| MoSold | 售出月份 |
| YrSold | 售出年份 |
| SaleType | 销售类型 |
| SaleCondition | 销售条件 |
| SalePrice | 房价（美元） |

### **实验步骤**

1. 数据导入：利用Pandas将房价数据集导入到DataFrame
2. 数据认识与探索：

* 使用pandas的info()方法探究属性的特点
* 找出所有的object类型字段，并使用value\_counts()查看具体原因
* 利用Pandas可视化MSZoning不同取值的记录数的柱状图
* 对预测的二手房价格进行可视化，查看是否有长尾效应
* 使用Pandas的isnull()方法查看有空缺值的属性，并使用fillna()方法用每个属性的众数对缺失值进行填充

1. 数据变换：

* 对object的所有标称属性进行类别编码
* 使用Pandas的apply方法对YearBuilt、YearRemodAdd两个日期字段计算距离2011年1月1日0时0分0秒的天数
* 使用Pandas的apply方法对需要归一化的字段进行归一化处理

### **小结**