数据科学导论实验 2：房价数据预处理

一、项目概述

本项目基于已有的房价数据集进行数据预处理，包括：
- 缺失值检测与处理
- 异常值检测与处理
- 特征间相关性分析
- 数据标准化与离散化

二、数据集信息

- 数据集：房价数据集 (train.csv 文件)
- 样本数量：1460 条
- 原始特征数量：81 个
- 目标变量：SalePrice

三、技术实现

1. 主要工具库
- pandas：数据处理和分析
- numpy：数值计算和数组操作
- scikit-learn：机器学习算法和预处理
- json：用于读取 json 文件
- seaborn：数据可视化
- matplotlib：图表绘制
- scipy：统计分析

2. 关键算法和函数
- KNNImputer：缺失值填充
- LinearRegression：异常值预测
- StandardScaler：数据标准化
- LabelEncoder：分类特征编码

四、文件结构

```
第二次作业/
├── main.ipynb                主程序文件
├── data/
│   ├── train.csv             原始数据集
│   └── data_description.txt  数据描述文件
│   └── feature_values.json   分类型特征值范围定义
└── README.pdf                项目说明文档
```

五、实验步骤

1. 缺失值检测与处理

1.1. 缺失值检测

- 计算每个特征的缺失值数量和比例
- 发现 19 个特征存在缺失值，占总特征的 23.5%
- 使用库和函数：pandas.isnull( ), pandas.sum( )

```
缺失值统计：
          特征名称  缺失数量     缺失比例(%)
72        PoolQC  1453  99.520548
74   MiscFeature  1406  96.301370
6          Alley  1369  93.767123
73         Fence  1179  80.753425
25    MasVnrType   872  59.726027
57   FireplaceQu   690  47.260274
3    LotFrontage   259  17.739726
58    GarageType    81   5.547945
59   GarageYrBlt    81   5.547945
60  GarageFinish    81   5.547945
63    GarageQual    81   5.547945
64    GarageCond    81   5.547945
35  BsmtFinType2    38   2.602740
32  BsmtExposure    38   2.602740
33  BsmtFinType1    37   2.534247
31      BsmtCond    37   2.534247
30      BsmtQual    37   2.534247
26    MasVnrArea     8   0.547945
42    Electrical     1   0.068493
有缺失值的特征数/总特征数：19/81
```

1.2. 缺失值处理

- 删除策略：删除缺失比例超过 40%的特征
  - 删除特征: ['PoolQC', 'MiscFeature', 'Alley', 'Fence', 'MasVnrType','FireplaceQu']
  - 使用库和函数：pandas.drop( )
- 填充策略:
  - 数值型特征：使用 KNN 模型预测缺失值 (n_neighbors=5)
    使用库和函数：sklearn.impute.KNNImputer
  - 分类型特征：使用众数填充
    使用库和函数：pandas.mode(), pandas.fillna( )

2. 异常值检测与处理

2.1. 分类型特征异常值处理

- AI 读取 data/data_description.txt，建立所有分类型特征的键值对文件 feature_values.json

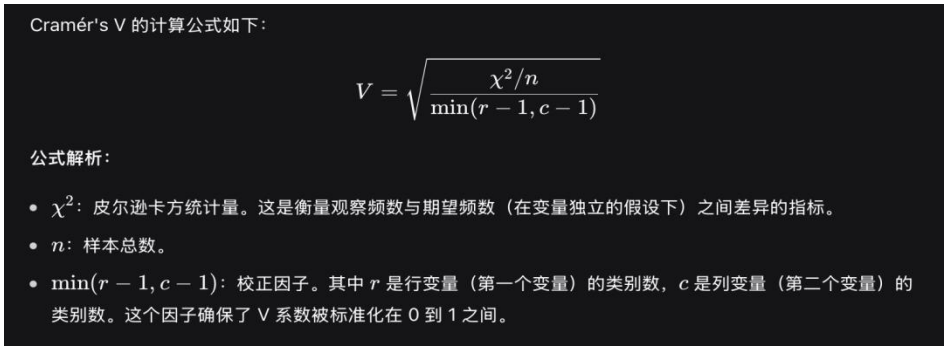
```
{
  "MSSubClass": ["20", "30", "40", "45", "50", "60", "70", "75", "80", "85", "90", "120", "150", "160", "180", "190"],
  "MSZoning": ["A", "C", "FV", "I", "RH", "RL", "RP", "RM"],
  "Street": ["Grvl", "Pave"],
  "Alley": ["Grvl", "Pave", "NA"],
  "LotShape": ["Reg", "IR1", "IR2", "IR3"],
  "LandContour": ["Lvl", "Bnk", "HLS", "Low"],
  "Utilities": ["AllPub", "NoSewr", "NoSeWa", "ELO"],
  "LotConfig": ["Inside", "Corner", "CulDSac", "FR2", "FR3"],
  "LandSlope": ["Gtl", "Mod", "Sev"],
```

- 基于 feature_values.json 中定义的有效值范围检测异常值
- 发现并修正的异常值：
  - MSZoning: 'C (all)' → 'C'
  - Neighborhood: 'NAmes' → 'Names'
  - BldgType: 'Duplex' → 'Duplx',
    '2fmCon' → '2FmCon',
    'Twnhs' → 'TwnhsE'/'TwnhsI'（按已有比例随机赋值）
  - Exterior2nd: 'CmentBd' → 'CemntBd',
    'Brk Cmn' → 'BrkComm',
    'Wd Shng' → 'WdShing'
- 验证异常值处理
- 使用库和函数：json.load( ), pandas.replace( ), numpy.random.choice( )

2.2. 数值型特征异常值处理
- 定义房地产相关的合理范围（本报告中标准由 AI 生成）
- 检测到异常值的特征：LotArea (11 个)、LotFrontage (2 个)、TotalBsmtSF (1 个)、1stFlrSF (7 个)
- 使用 sklearn 中的线性回归模型预测异常值的真实值，并限制在合理范围内
- 验证异常值处理
- 使用库和函数：sklearn.linear_model.LinearRegression, numpy.clip( )

3. 特征相关性分析与冗余特征删除

3.1. 分类型特征冗余检测
- 使用 Cramér's V 系数分析分类型特征相关性


Cramér's V 的计算公式如下：

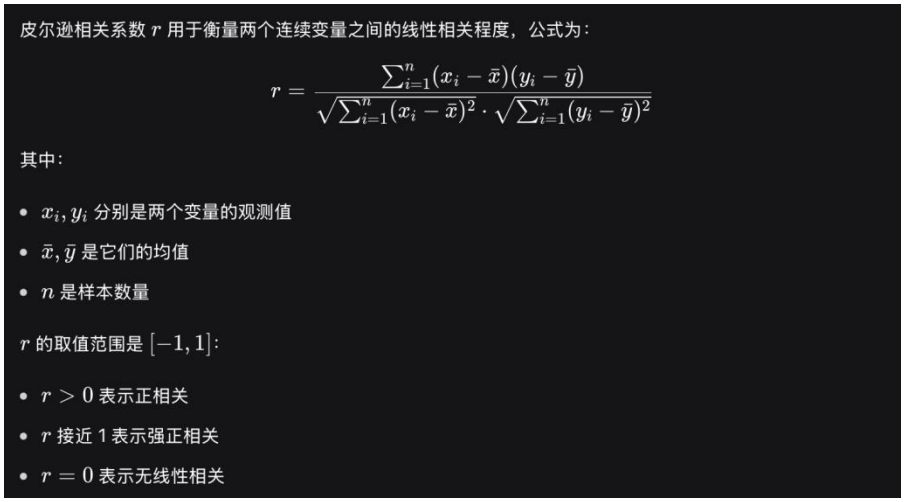
$$V = \sqrt{\frac{\chi^2/n}{\min(r-1, c-1)}}$$

公式解析：

- χ^2：皮尔逊卡方统计量。这是衡量观察频数与期望频数（在变量独立的假设下）之间差异的指标。
- n：样本总数。
- $\min(r-1, c-1)$：校正因子。其中 r 是行变量（第一个变量）的类别数，c 是列变量（第二个变量）的类别数。这个因子确保了 V 系数被标准化在 0 到 1 之间。

- 阈值设置：0.6，大于等于 0.6 的特征值视作冗余特征对
- 发现冗余特征对：
  - ('MSZoning', 'Neighborhood')
  - ('Exterior1st', 'Exterior2nd')
  - ('GarageQual', 'GarageCond')
- 删除冗余特征对中与 SalePrice 相关性较小的那个特征
  删除：['Neighborhood', 'Exterior2nd', 'GarageCond']
- 使用库和函数：
  - sklearn.preprocessing.LabelEncoder：将特征转换为数字编码，便于后续分析。
  - scipy.stats.chi2_contingency：用于卡方检验，衡量两个分类变量之间的相关性。
  - pandas.crosstab( )：生成分类变量的列联表，用于统计频数关系。

3.2. 数值型特征冗余检测



- 使用皮尔逊相关系数分析数值型特征相关性
- 阈值设置：0.8，绝对值大于等于 0.8 的特征值视作冗余特征对
- 发现冗余特征对：
  - ('GrLivArea', 'TotRmsAbvGrd')
  - ('TotalBsmtSF', '1stFlrSF')
  - ('GarageCars', 'GarageArea')
- 删除冗余特征对中与 SalePrice 相关性较小的那个特征
  删除：['GarageArea', 'TotRmsAbvGrd', '1stFlrSF']
- 使用库和函数：pandas.corr(method='pearson')：皮尔逊相关系数计算相关性

4. 目标变量标准化与离散化

4.1. 标准化处理
- 对 SalePrice 进行 Z-score 标准化，将数据转换为均值为 0，标准差为 1 的标准正态分布。

  公式：

$$z = \frac{x - \mu}{\sigma}$$

  其中，
  x: 原始数据
  μ: 该特征的均值
  σ: 该特征的标准差
  z: 标准化后的数据

- 生成新特征：SalePrice_std
- 使用库和函数：sklearn.preprocessing.StandardScaler

4.2. 离散化处理
- 等宽离散化：将 SalePrice 分为 5 个等宽区间（每个分箱长度相等）
  - 分界线：[34180, 178920, 322940, 466960, 610980, 755000]

  - 使用库和函数：pandas.cut( )
- 等频离散化：将 SalePrice 分为 5 个等频区间（每个分箱样本数量大致相等）
  - 分界线：[34900, 124000, 147000, 179280, 230000, 755000]
  - 使用库和函数：pandas.qcut( )
- 两种离散方式的合理性分析

由于 SalePrice 区间样本量失衡，低价区间包含大量样本，高价区价格跨度极大且样本极少，等宽离散化易导致存在样本量极少甚至为空的分箱。

等频离散化确保每个区间包含大致相同数量的样本，每个区间都有足够的样本量进行分析，统计稳定性良好。

5. 与 SalePrice 相关性最高的三个特征分析

- 与 SalePrice 相关性皮尔逊相关性最高的三个特征：

・No1. GrLivArea (0.708624)：地面以上居住面积（平方英尺）

房屋价格=房屋单价×居住面积，虽然房屋单价因地价、环境、交通等因素存在不同，但房屋价格与居住面积存在强烈的正相关关系。

・No2. GarageCars (0.640409)：车库容量（可容纳车辆数量）

在欧美住宅市场，车库不仅用于停车，还可以作为储藏室或工作间。能容纳多辆车的车库往往对应较高价值的住宅，因此也与房价显著相关。

・No3. TotalBsmtSF (0.633785)：地下室总面积（平方英尺）

地下室面积一般与房屋面积存在正相关关系，故与房屋价格显著相关符合直觉。

- 使用库和函数：pandas.corr(method='pearson')