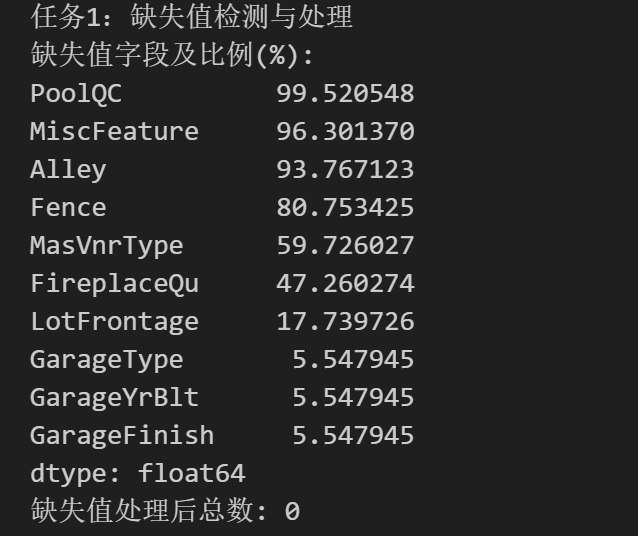
本次任务较为复杂，以下是我的理解与阐述。

**任务1：缺失值的检测与缺失值处理**

我先用pandas读取train数据集中的数据，后打印出数据集的形状，通过isnull()的求和判断缺失值总数



计算各字段缺失值比例并输出，按降序排序，展示前 10 个缺失率最高的字段，直观呈现数据缺失情况。



对于分类变量，将缺失值填充为'None'，表示无该特征，符合实际意义。

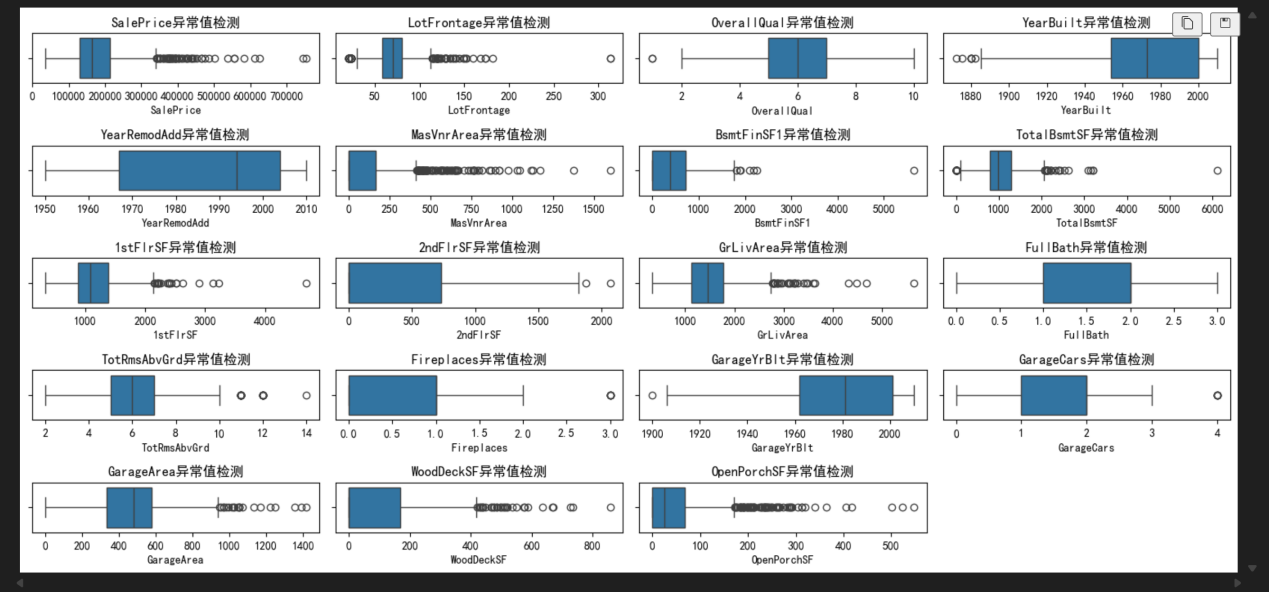
针对数值变量，选取缺失的数值列，使用KNNImputer进行填充，n\_neighbors=5表示利用最近的 5 个邻居样本数据进行填充，使填充值更具合理性。

输出处理后的缺失值总数为0，验证处理效果良好。

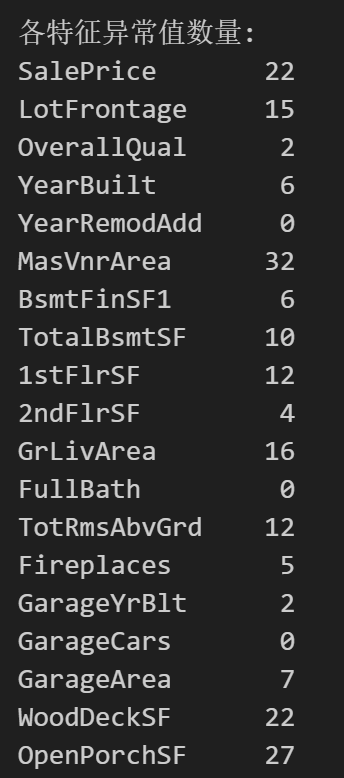
**任务2：异常值检测**

在众多信息列中，我选取与房价相关性大于 0.3 的数值特征和房价特征一起作为核心特征，用于异常值检测，确保检测的针对性和有效性。（用于异常值检测的核心特征: ['SalePrice', 'LotFrontage', 'OverallQual', 'YearBuilt', 'YearRemodAdd', 'MasVnrArea', 'BsmtFinSF1', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'GrLivArea', 'FullBath', 'TotRmsAbvGrd', 'Fireplaces', 'GarageYrBlt', 'GarageCars', 'GarageArea', 'WoodDeckSF', 'OpenPorchSF']）

使用箱线图可视化各核心特征的分布，直观展示异常值。如图所示，蓝色方框为箱形部分，代表数据的四分位区间，方框的下沿是数据的下四分位数（Q1），表示有 25% 的数据小于该值，方框的上沿是数据的上四分位数（Q3），表示有 25% 的数据大于该值，方框中间的线：是数据的中位数（Q2），表示有 50% 的数据小于该值。圆圈代表异常值，小于 Q1 - 1.5×IQR 或 大于 Q3 + 1.5×IQR 的数据，会被标记为异常值，这里的 IQR = Q3 - Q1（上四分位数减下四分位数，即中间 50% 数据的范围），用于直观展示数据中“偏离大多数样本的极端值”，即任务二要求我们检测的异常值。



依据 3σ 法则计算各特征的异常值范围，统计异常值数量，了解异常值分布情况。



采用分位数截断法处理异常值，将小于 5% 分位数的值设为 5% 分位数，大于 95% 分位数的值设为 95% 分位数，避免异常值对后续分析的干扰。

**任务三：特征间的相关性分析**

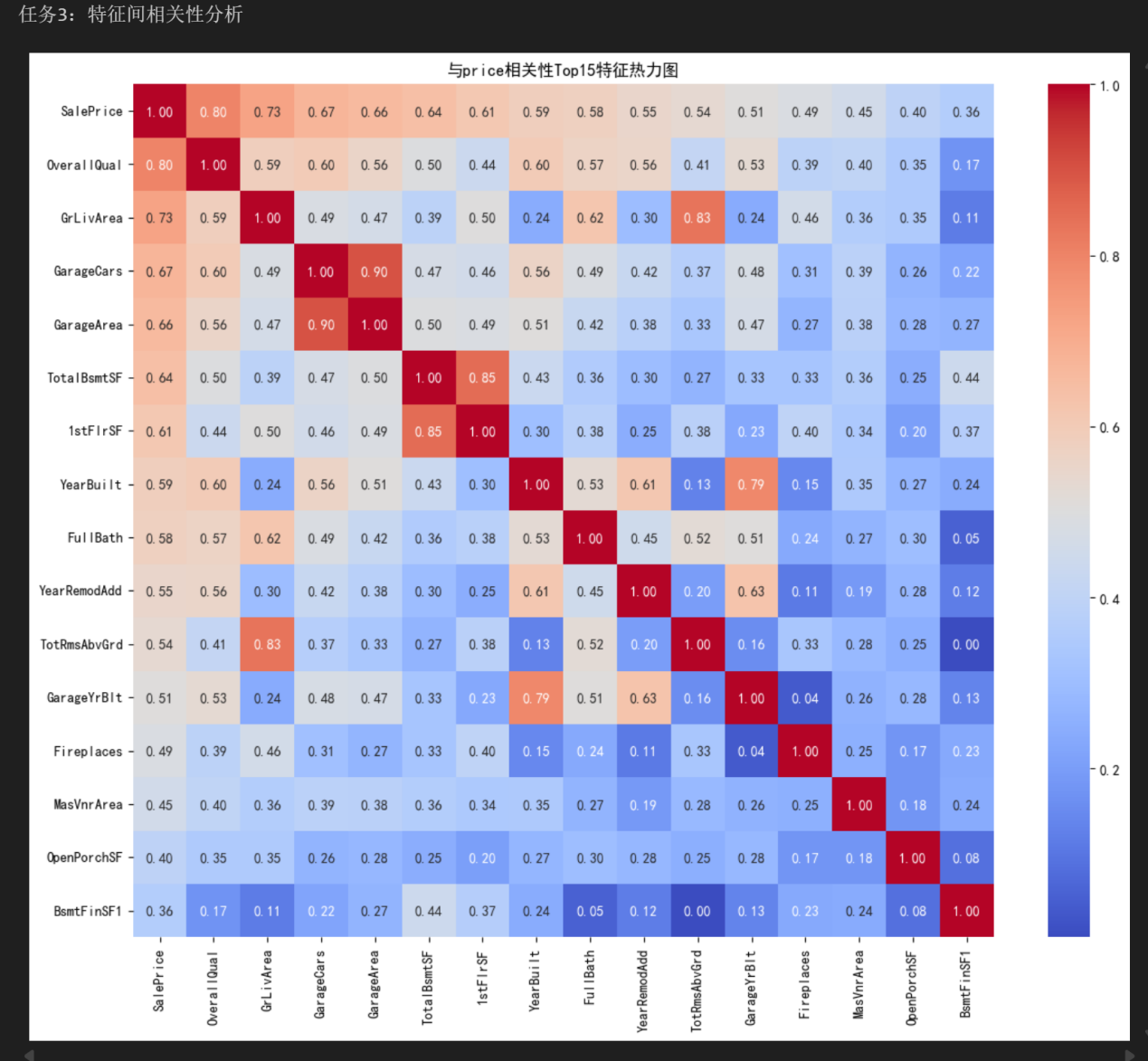
通过df.select\_dtypes(include=[np.number])筛选出所有数值类型的列，过滤掉文本和分类变量，用.corr() 计算这些数值列之间的Pearson相关系数，得到一个正方形的相关性矩阵 corr\_matrix。矩阵中每个元素 corr\_matrix[i, j] 表示第 i 列和第 j 列的相关系数（范围 [-1, 1]，越接近 1 正相关越强，越接近 -1 负相关越强）。

对所有相关系数取绝对值，忽略掉负相关和正相关因素，只在意相关程度大小进行排序，选取与房价相关性绝对值排名前 15 的特征，绘制热力图，通过颜色直观展示特征间相关性的强弱，便于发现强相关和弱相关特征。

热力图如图所示，每个单元格的颜色和数值共同展示了两个特征之间的相关程度。比如：

第一行第一列（SalePrice 与自身）的相关系数是 1.00（完全正相关）。OverallQual与 SalePrice 的相关系数是 0.80（强正相关，说明房屋质量越高，房价通常越高）。

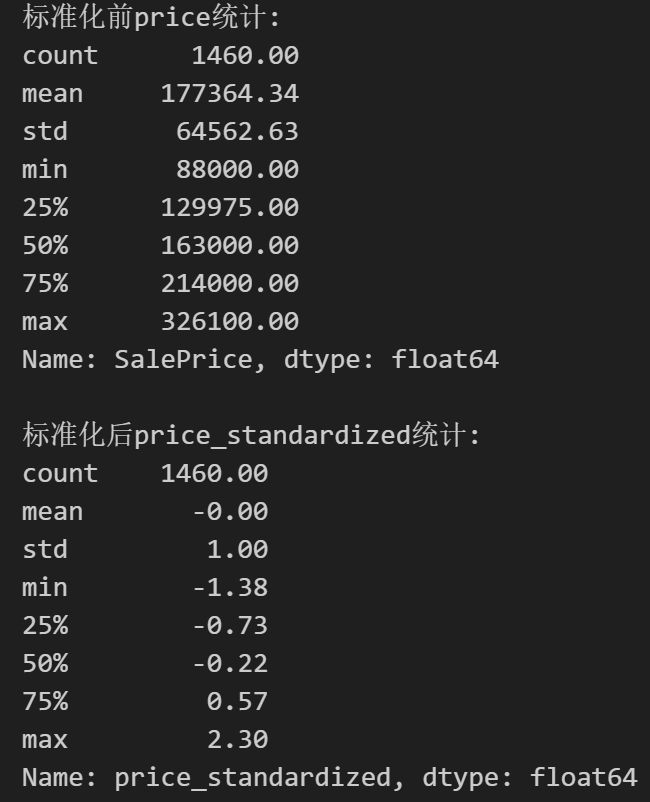
颜色越红，正相关越强；颜色越蓝，负相关越强；颜色越浅，相关程度越弱。



**任务四：对price属性进行标准化**

通过scaler.fit\_transform(df[['SalePrice']])对房价（SalePrice）列进行标准化处理，并将结果存储在新列price\_standardized中。fit\_transform方法先拟合数据（计算均值和标准差），再进行转换。

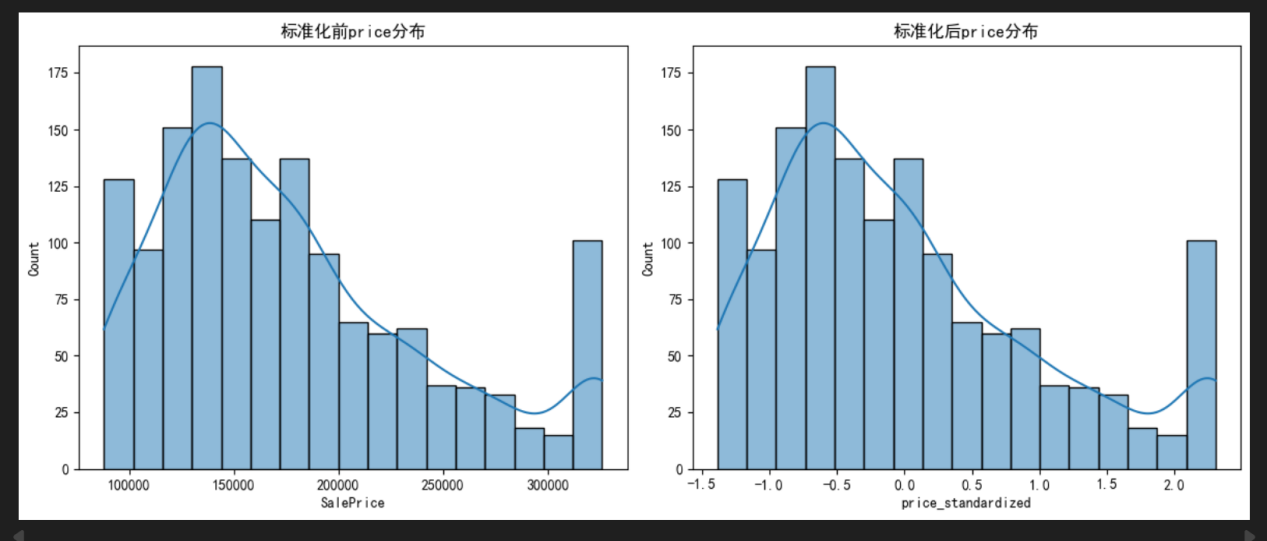
输出标准化前房价（SalePrice）的统计信息，包括均值、标准差、最小值、最大值等并保留两位小数，同样输出标准化后房价（price\_standardized）的统计信息，对比可发现标准化后均值接近 0，标准差接近 1，符合标准正态分布的特征。



后续构造两个子图进行分布可视化，左侧子图用seaborn的histplot绘制标准化前房价的直方图，并结合kde=True添加核密度估计曲线，直观展示原始房价的分布情况。

右侧子图绘制标准化后房价的直方图和核密度估计曲线。

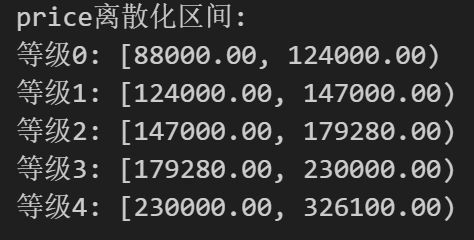
如图所示，对比可看出，标准化后数据分布更接近正态分布，且在 0 附近集中。标准化前 SalePrice：mean=177364.34，std=64562.63；标准化后 price\_standardized：mean近似于0，std=1.00，成功完成了均值归 0，标准差归 1的标准化目标



**任务五；根据price属性进行离散化**

先进行离散化器初始化，使用KBinsDiscretizer类初始化离散化器discretizer。n\_bins=5表示将数据分为 5 个区间，并将分箱结果编码为顺序值，表示采用分位数策略进行分箱，指定计算分位数的方法为线性。

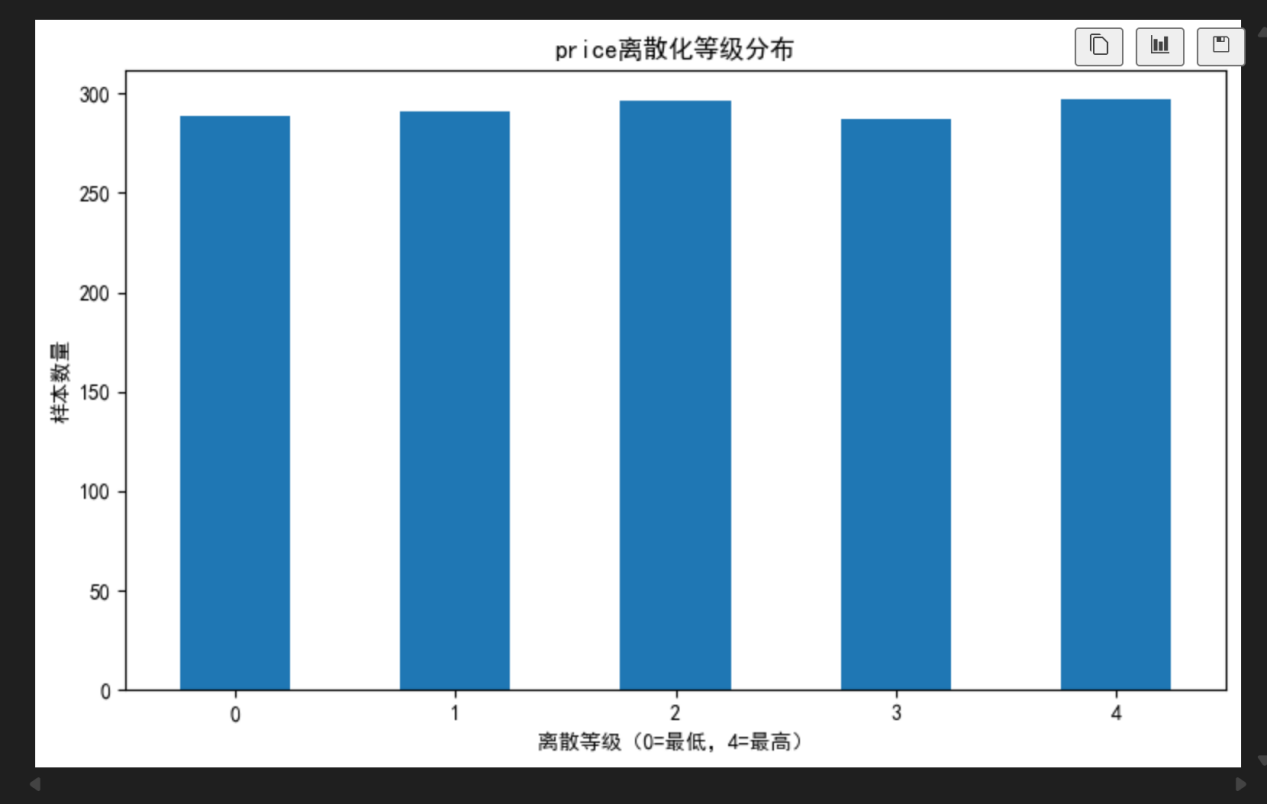
通过discretizer.fit\_transform(df[['SalePrice']])对房价列进行离散化处理，将结果转换为整数类型后存储在新列price\_discretized中。获取离散化的区间边界bin\_edges，循环输出每个离散等级对应的房价区间。



统计每个离散等级的样本数量，并按等级排序输出。

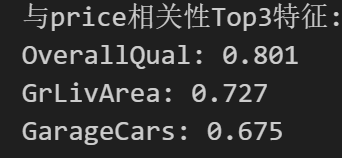


绘制柱状图展示各离散等级的样本分布情况，如图所示。

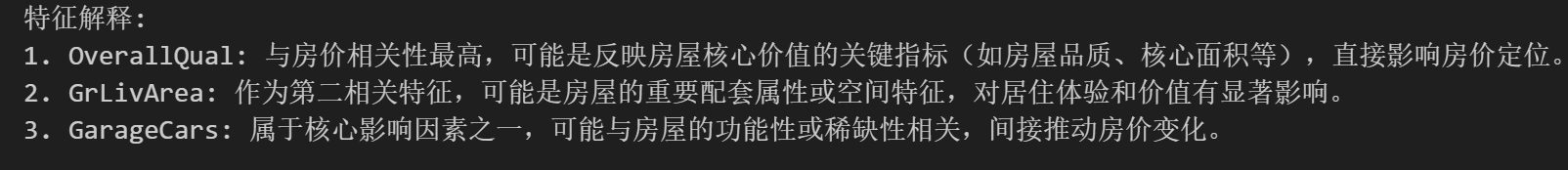
。

**任务六：找出与price相关性最高的三个特征及解释**

从任务三计算得到的相关系数矩阵corr\_matrix中，提取SalePrice列的相关系数，然后按降序排序，得到price\_corr。由于房价与自身的相关性为 1.0，所以选取排序后第 2 到第 4 个特征，即与房价相关性最高的三个特征，并循环输出。



对三个特征进行解释，从房屋品质、配套属性、功能性和稀缺性等方面推测它们影响房价的原因，为后续房价分析提供参考。



**小总结：**通过此次实验我接触到了标准化和离散化对于数据集及后续相关分析的重要性，标准化使数据具有统一尺度，便于与其他特征综合分析；离散化将数据分类，更清晰地呈现数据分布特征。也学会了热力图的绘制方法，通过颜色深浅呈现与主体相关系数的强弱程度。