房价预测分析报告

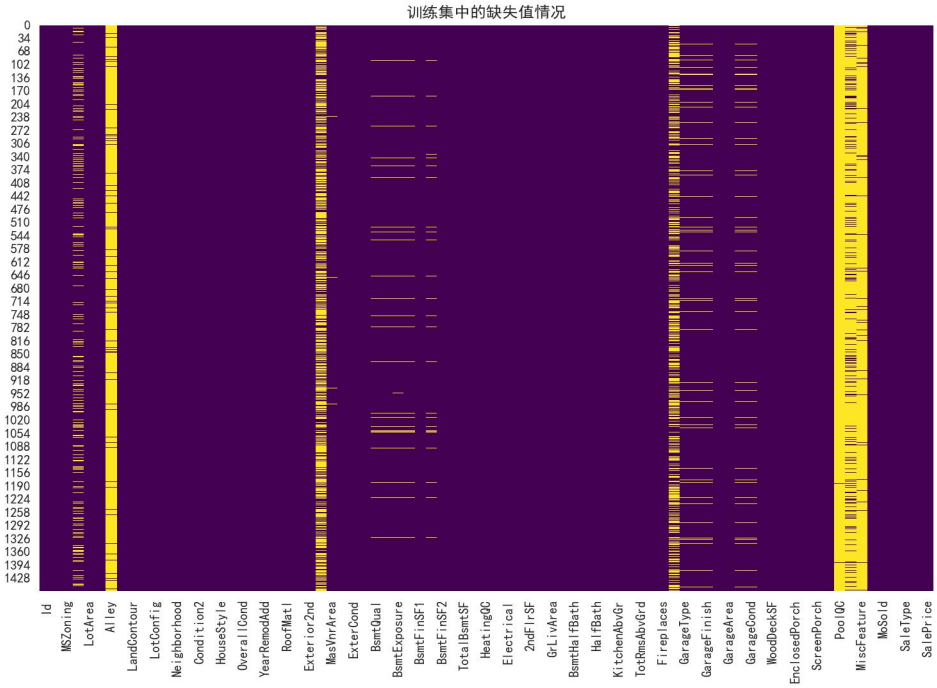
1 数据概览

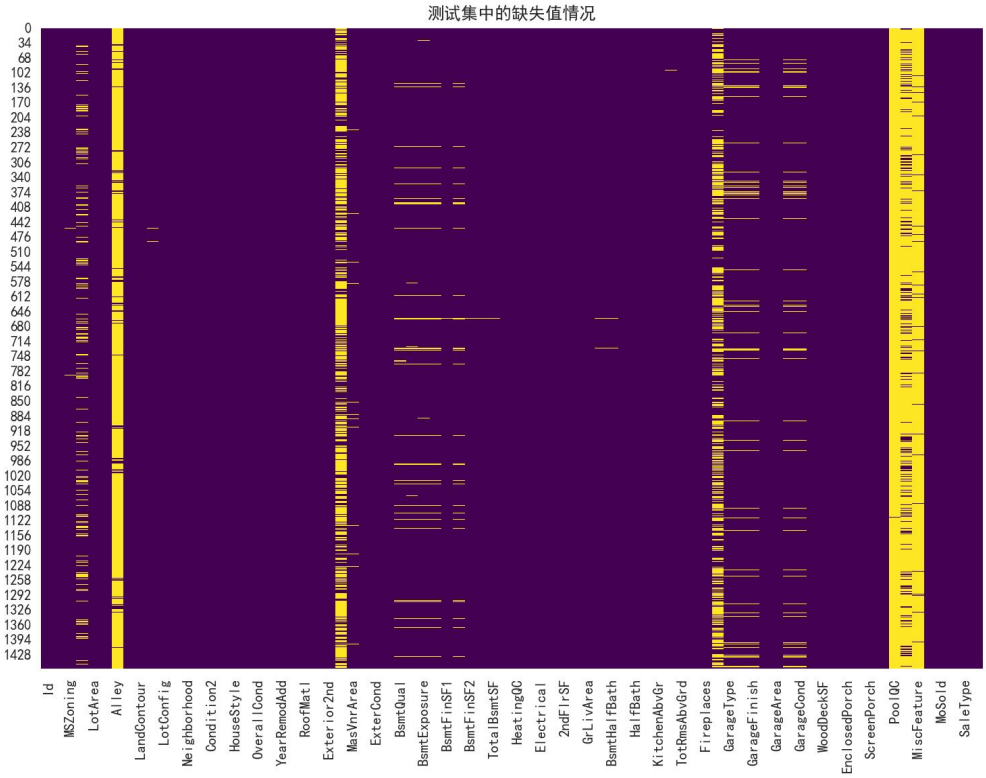
本次分析使用的数据集分为训练集和测试集。训练集包含 1460 条记录和 81 个特征，而测试集包含 1459 条记录和 80 个特征。训练集中的额外特征 `SalePrice` 是目标变量，用于表示房屋价格。

数据集中的特征可以大致分为数值特征和类别特征，涵盖了房屋的各个方面，如房屋的整体质量(`OverallQual`)、居住面积(`GrLivArea`)、车库容量(`GarageCars`)等。训练集和测试集的基本结构提供了全面的信息，方便进行更深入的分析。

2 缺失值分析

通过对训练集和测试集进行缺失值分析，发现一些特征存在较多的缺失值。以下是缺失值可视化结果：





在训练集中，PoolQC（游泳池质量）、MiscFeature（其他附加功能）、Alley（巷道类型）等特征的缺失值比例较高。这些特征的缺失值可能是因为并非所有房屋都有这些特征，从而导致数据中缺失值较多。

在处理这些缺失值时，对于缺失值比例高且不重要的特征，可以考虑删除该特征。而对于重要性较高的特征，可以使用均值、中位数或最频繁值进行填充，具体选择应根据特征的分布和业务需求来决定。对于类别特征，可以使用众数填充或引入新的类别“缺失”来处理缺失值。

3 数据预处理

3.1 初步探索

• **数据概览**：首先进行数据的初步统计分析，了解各特征的分布、缺失值情况及潜在的异常值。

• **特征类型确认**：区分数值型与类别型特征，为后续处理做准备。

3.2 数据清洗

• **缺失值处理**：对于含有缺失值的记录，根据特征的性质，采用合适的策略处理，如均值填充对于连续数值特征。

• **异常值检测与处理**：利用箱线图等方法识别并适当处理数据集中的极端值。

3.3 数值特征处理

• **偏态修正**：对于偏斜分布的数值特征，采用对数变换 $X' = \log(X + 1)$，以减少长尾效应，使分布更接近正态分布。

3.4 类别特征编码

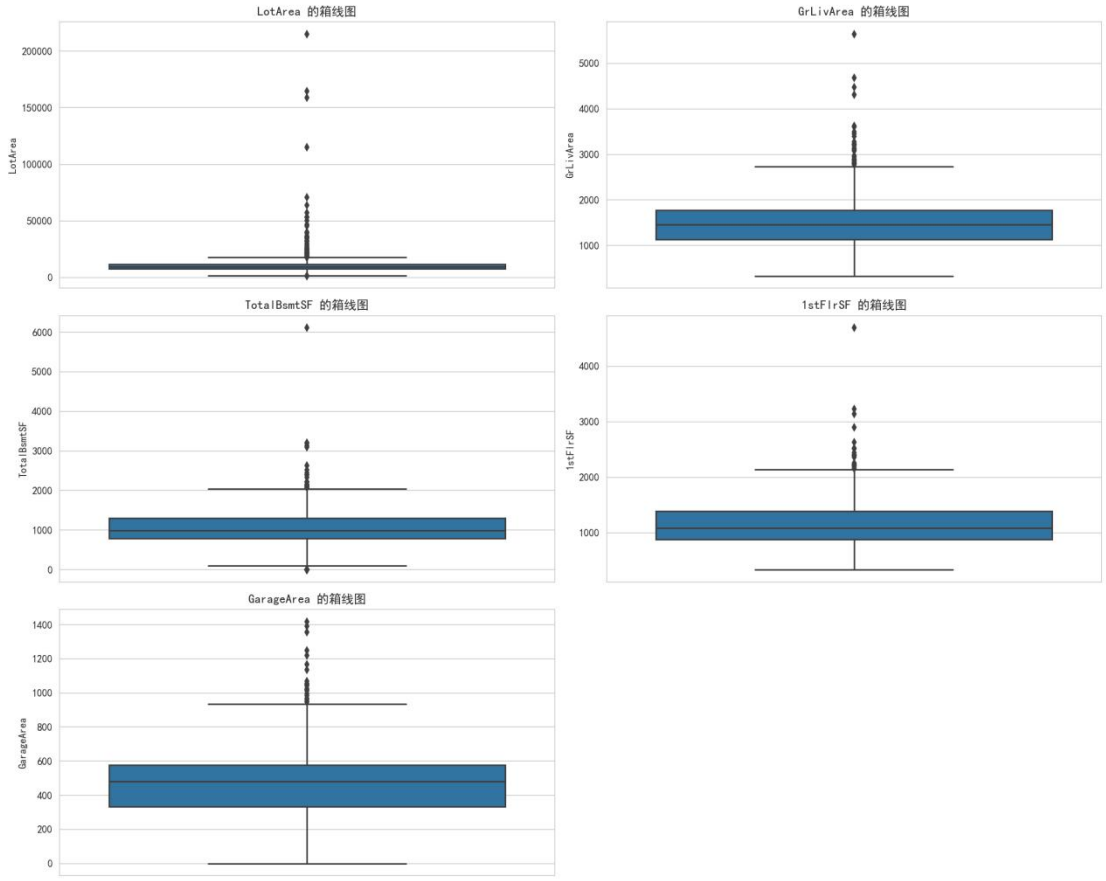
• **One-Hot 编码**：将类别特征转换为一系列二进制特征，便于模型理解和计算。

3.5 特征标准化

• **标准化处理**：利用 StandardScaler，将数值特征缩放至同一尺度，公式为 $Z = (X - \mu) / \sigma$，其中μ是均值，σ是标准差。

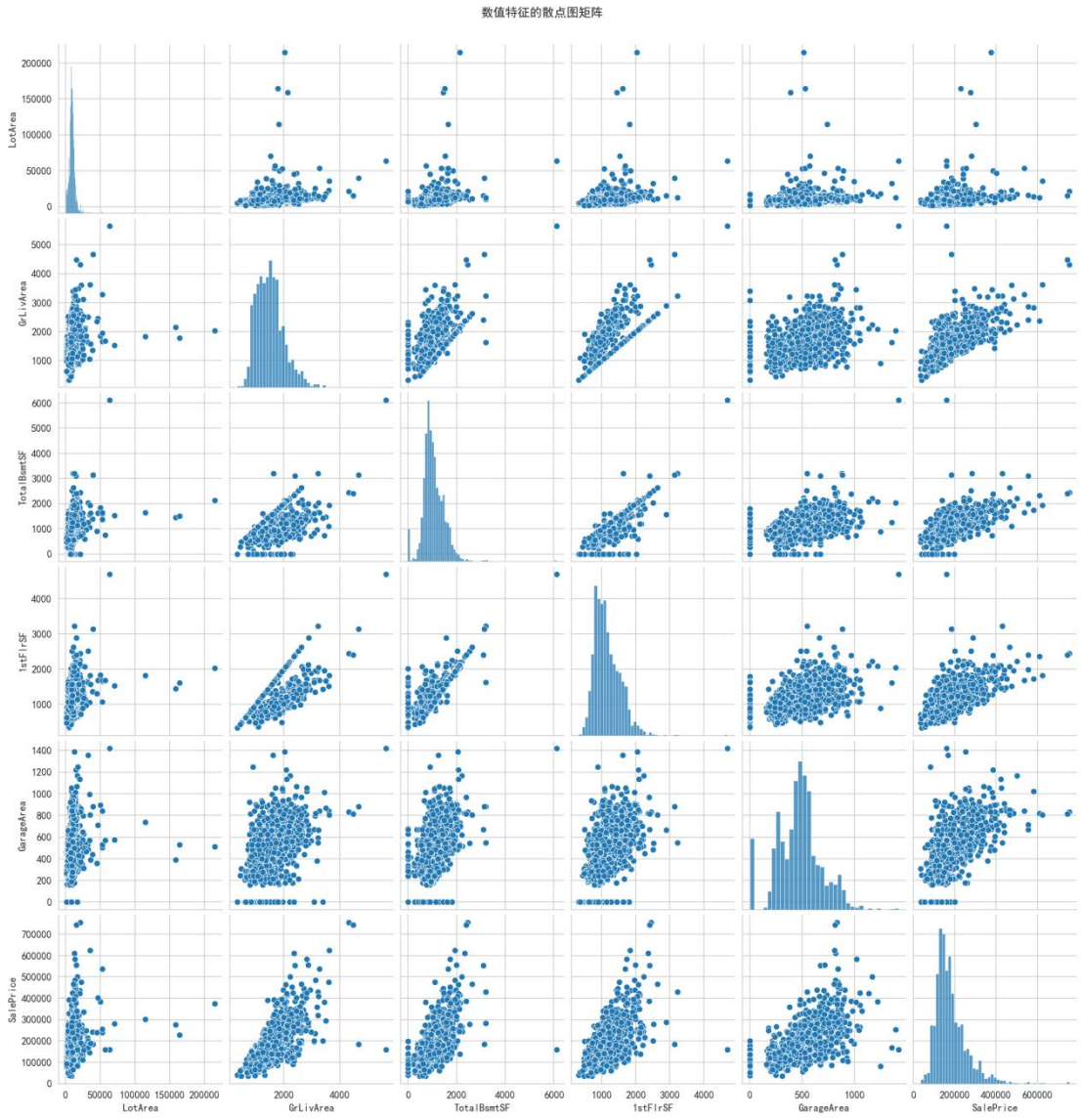
4 数值特征分析

本文选择了一些重要的数值特征进行详细分析，包括 LotArea（地块面积）、GrLivArea（居住面积）、TotalBsmtSF（地下室面积）、1stFlrSF（一楼面积）和 GarageArea（车库面积）。以下是这些特征的箱线图：



从箱线图中可以看出，LotArea 和 GrLivArea 存在较多离群值，这些离群值可能会影响模型的性能。而 TotalBsmtSF、1stFlrSF 和 GarageArea 的分布较为集中，但也存在一些离群值，这些离群值需要进一步处理以提升模型性能。

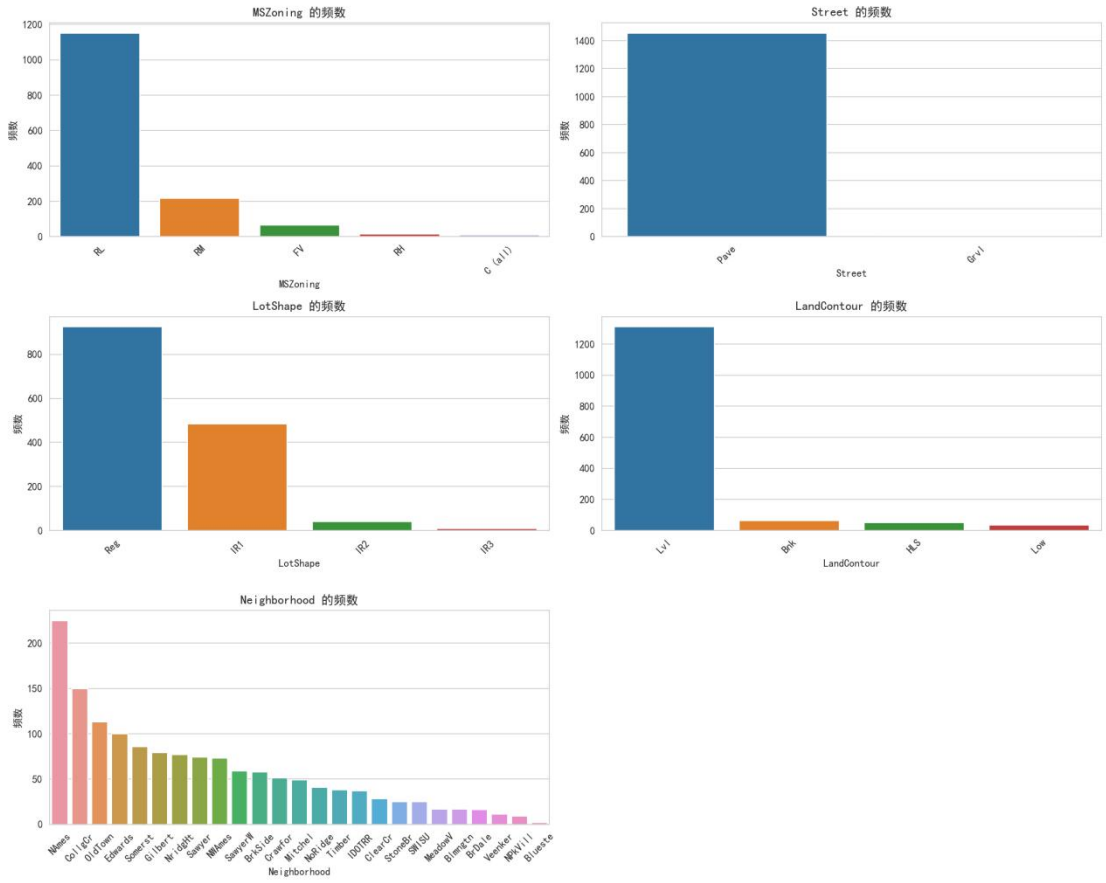
此外，通过数值特征的散点图矩阵，可以观察到各数值特征之间的关系以及与目标变量 SalePrice 的关系：



从散点图矩阵可以看出，部分特征之间存在较强的线性关系，如 GrLivArea 和 TotalBsmtSF。目标变量 SalePrice 与数值特征之间也存在一定的相关性，这为进一步的分析提供了基础。

5 类别特征分析

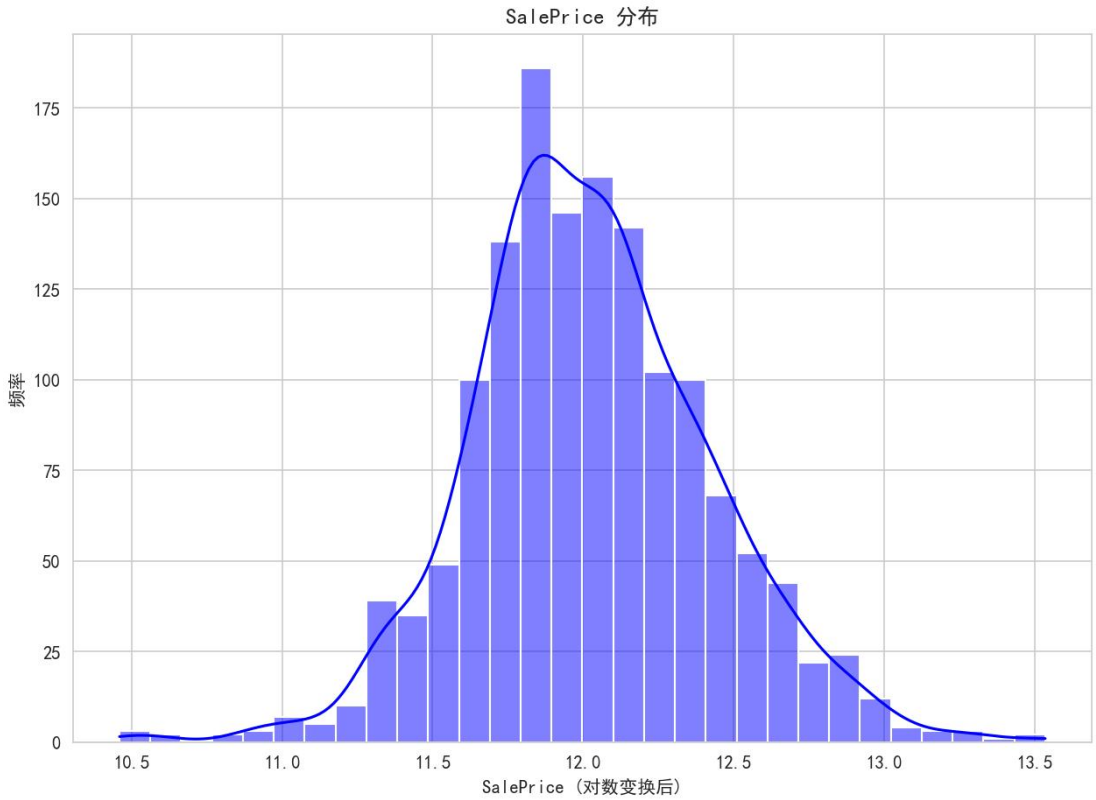
本文选取了几个重要的类别特征进行分析，包括 MSZoning（分类区划）、Street（街道类型）、LotShape（地块形状）、LandContour（土地轮廓）和 Neighborhood（邻居情况）。以下是这些特征的条形图：



通过条形图可以看出，MSZoning 以 RL（住宅用地）为主，Street 中 Pave（铺设道路）占绝大多数，而 LotShape 和 LandContour 也有较明显的类别分布。Neighborhood 中特定社区如 NAmes、CollgCr 的房屋数量较多。这些类别特征的频数分布为后续的分析提供了参考，帮助了解数据的总体分布情况。

6 目标变量分析

SalePrice 是要预测的目标变量，其分布情况如下：



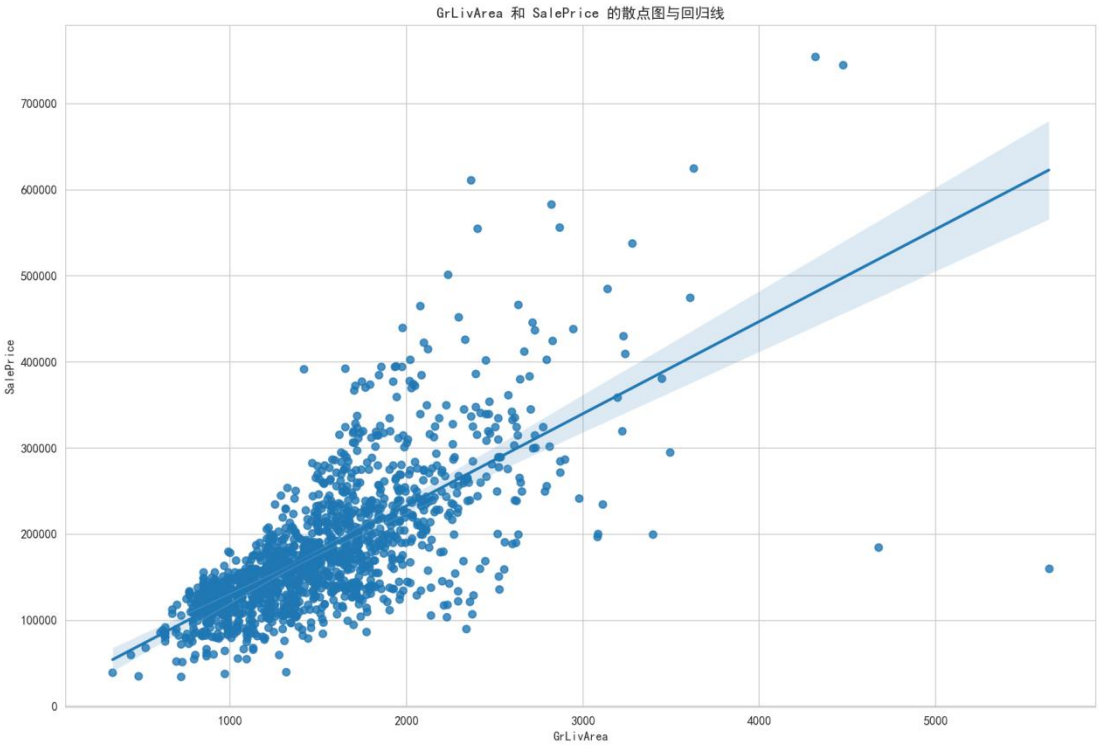
从分布图中可以看出，SalePrice 的分布呈右偏态，表明大部分房屋价格集中在较低的范围内。具体统计描述如下：

均值为 180921.195，中位数为 163000.0，标准差为 79442.502。这些统计描述为了解目标变量的基本特征提供了数据支持。

7 特征与目标变量的关系

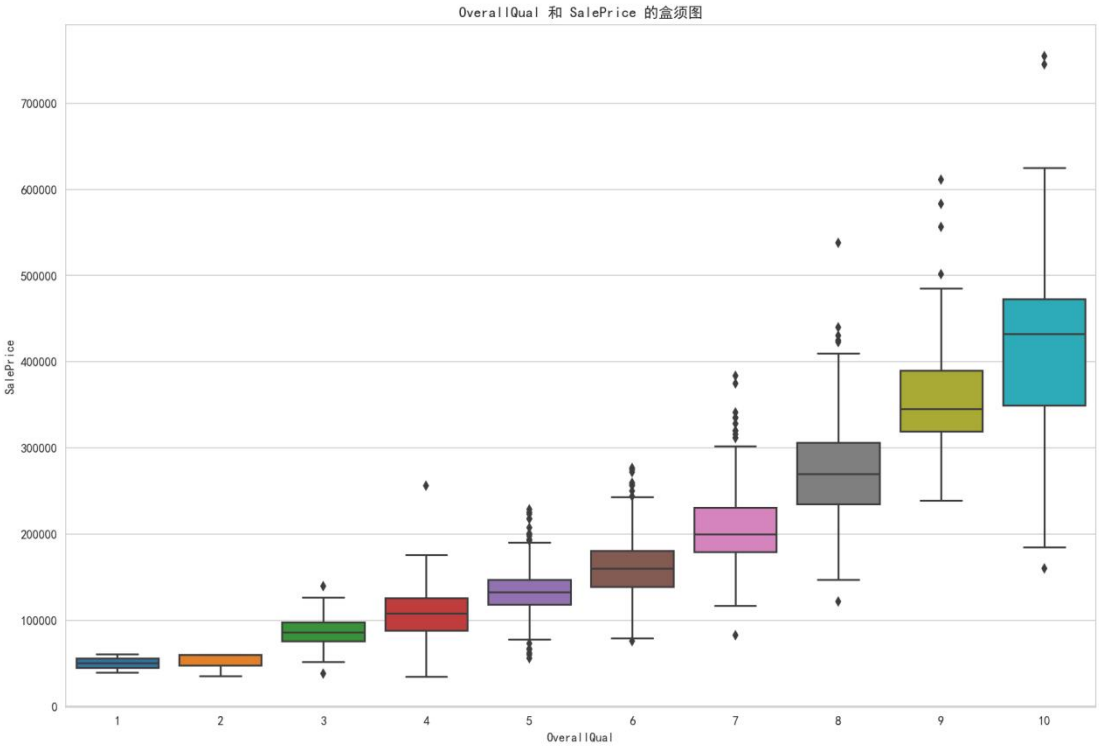
本文分析了数值特征与 SalePrice 的相关性，结果如下：

GrLivArea 与 SalePrice 的散点图与回归线：



可以看出，GrLivArea 与 SalePrice 存在较强的正相关关系，居住面积越大，房价越高。这说明居住面积是影响房价的重要因素。

OverallQual 和 SalePrice 的盒须图：



从图中可以看出，房屋的整体质量越高，房价也越高。这进一步证明了房屋质量在房价预测中的重要性。此外，GarageCars 也是一个重要特征，与 SalePrice 呈正相关关系。这些图表分析帮助识别出对房价影响较大的特征，为模型的特征选择和优化提供依据。

8 t-SNE 方法详解

8.1 t-SNE 基本原理

高维空间相似度：基于高斯分布计算点间相似度，即

$$p_{ij} = \frac{p_i p_j}{2\sigma_i^2} e^{-\frac{||x_i - x_j||^2}{2\sigma_i^2}} \qquad (1)$$

其中σ_i通过 perplexity 参数调整。

低维空间相似度：采用 t-分布作为相似度度量，即

$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq l} (1 + ||y_k - y_l||^2)^{-1}} \qquad (2)$$

8.2 t-SNE 目标函数

KL 散度：最小化高维空间中点对之间的相似度分布P与低维嵌入空间中点对相似度分布Q之间的 Kullback-Leibler 散度，即

$$\sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \qquad (3)$$

9 t-SNE 降维过程

9.1 初始化

随机设置低维空间中点的位置。

9.2 迭代优化

通过梯度下降法，不断更新低维空间中点的位置，直至达到收敛条件，最小化 KL 散度。

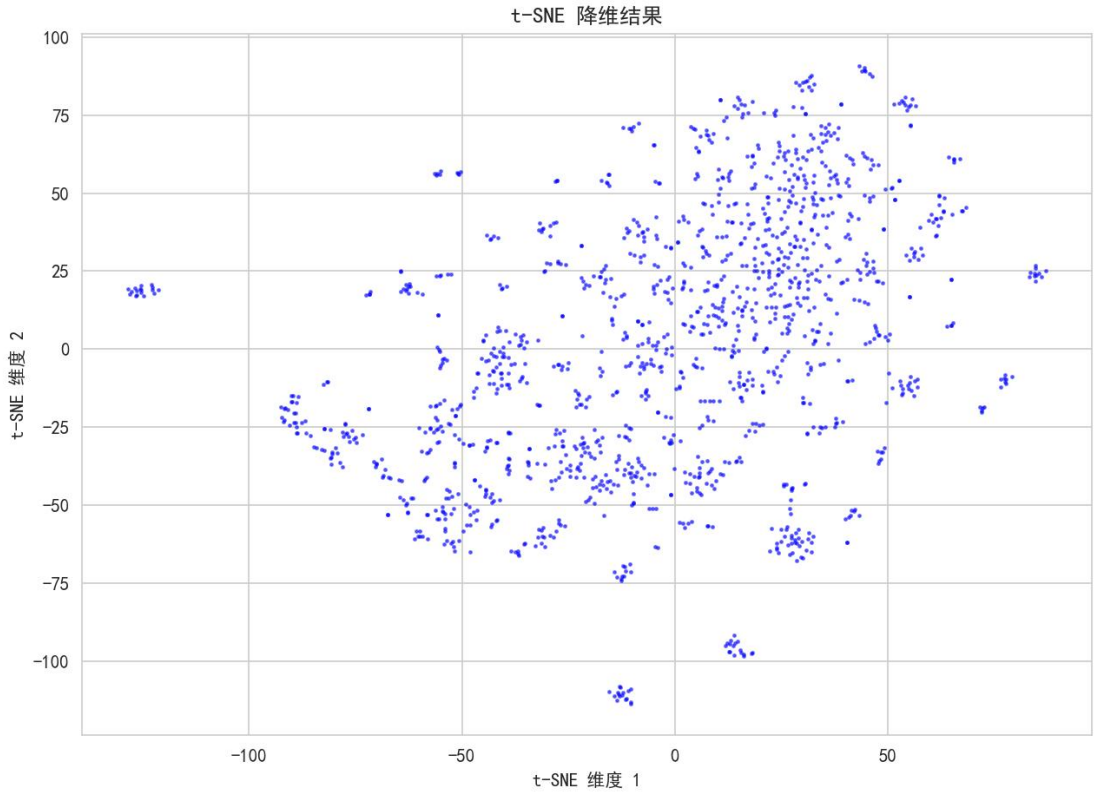
9.3 参数调整

Perplexity：控制了每个点的邻域大小，影响降维结果的聚类效果，通常需通过多次实验确定最优值。

10 t-SNE 降维结果分析

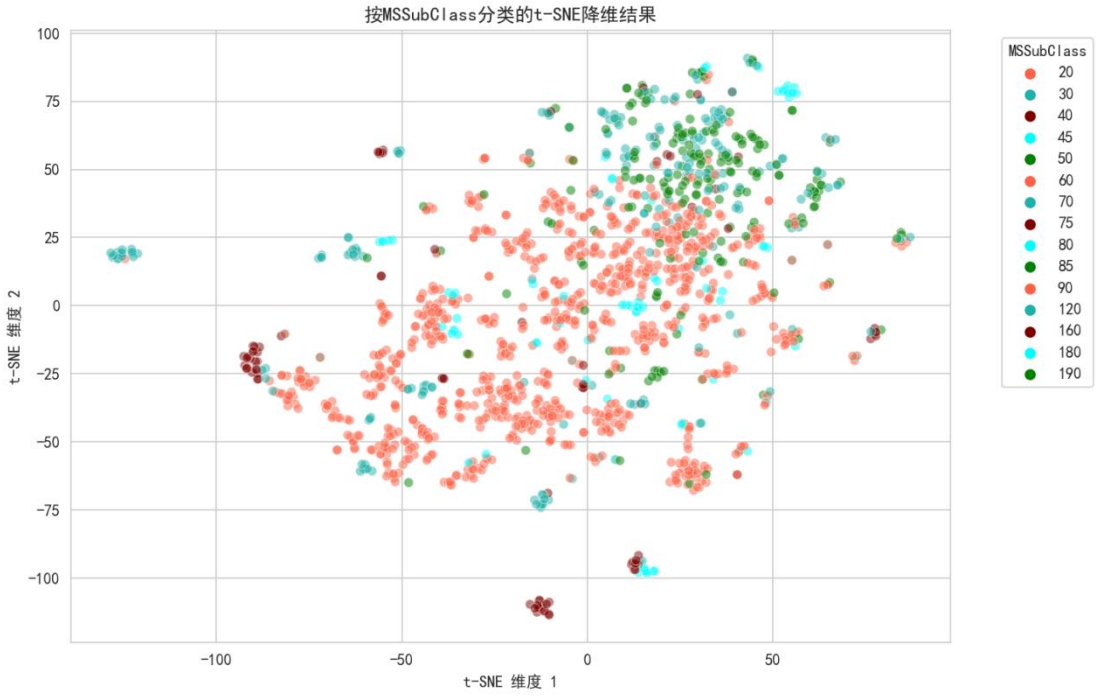
10.1 全局视角

• 生成的二维 t-SNE 图展现了数据在降维空间中的分布，有助于直观理解特征间的相互关系及潜在群组结构。



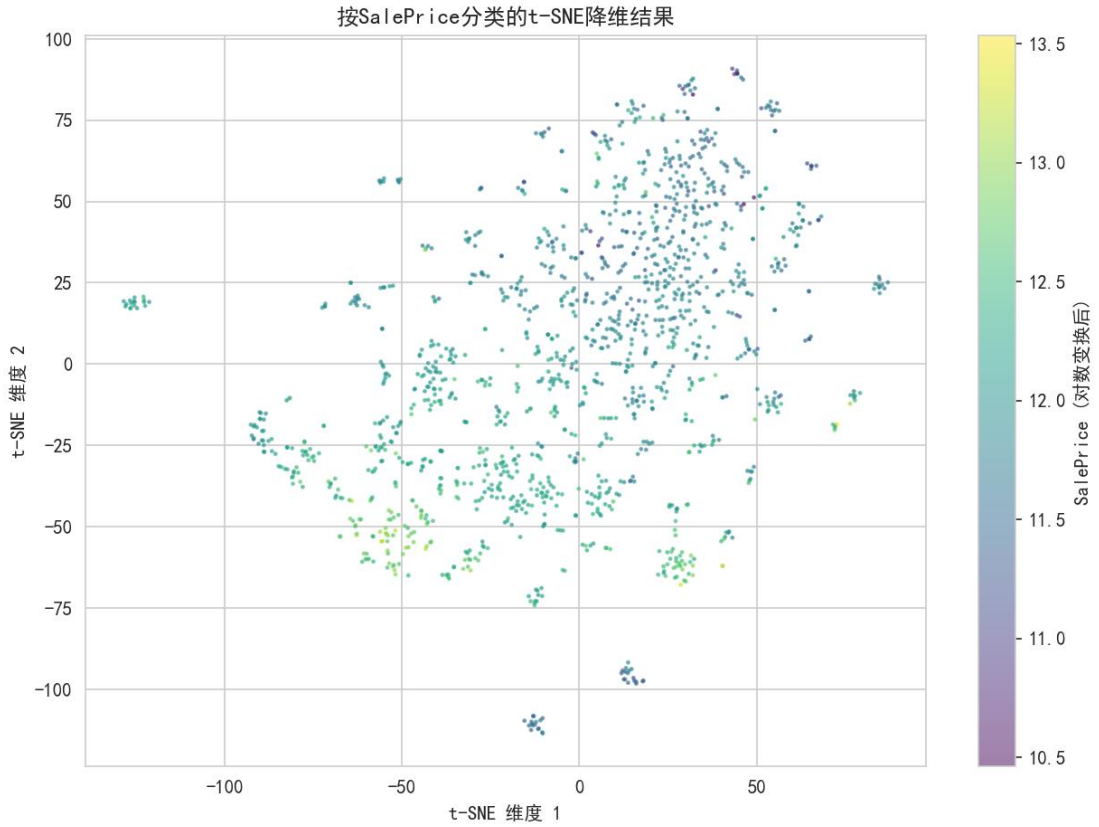
10.2 按类别特征分析

• 通过着色区分不同类别特征下的点群，可以观察到类别间明显的分界，显示 t-SNE 有效捕获了类别差异。



10.3 按目标变量分析

• 将目标变量（如对数变换后的房价）映射为颜色渐变，揭示了价格与特征空间位置的潜在相关性。



10.4 初步结论

通过对数据集的详细分析，本文得出了以下结论：

首先，数据集中存在一些缺失值较多的特征，需要进行适当处理。对于缺失值比例高且不重要的特征，可以考虑删除，而对于重要性较高的特征，建议使用填充方法。其次，数值特征中存在离群值，特别是 LotArea 和 GrLivArea，需要进一步处理以提升模型性能。此外，类别特征的分布情况多样，对于某些特征如 MSZoning 和 Neighborhood，可以进一步细化分析。

t-SNE 有效地将高维度的房地产数据降至二维，不仅保留了数据的重要结构信息，还通过可视化揭示了数据的内在模式。尽管 t-SNE 在探索性分析中表现卓越，但其计算成本和对超参数的选择敏感性要求在大规模数据应用中需谨慎考量。此外，t-SNE 更多用于数据探索，而不直接用于模型的输入层，因其不保证全局保距和保序性。

SalePrice 作为目标变量，呈现右偏分布，且与一些重要特征如 GrLivArea、OverallQual、GarageCars 等有较强的相关性。未来的房价预测模型应重点关注与 SalePrice 高度相关的特征，如 GrLivArea、OverallQual、GarageCars 等，进行特征选择和优化。这些分析为房地产市场的相关研究提供了数据支持，有助于相关方做出更科学的决策。