**Lab2：数据预处理的基本方法**

姓名：\_\_\_\_\_\_徐卓尔\_\_\_\_ 学号：\_\_\_\_10245102419\_\_\_\_\_\_

日期：\_\_\_\_9.22\_\_\_\_\_\_

# 一、实验目的

理解数据分析的基本流程，掌握缺失值处理、异常值检测、特征相关性分析、标准化与离散化等常见数据预处理方法。

# 二、实验数据与环境

数据集：train.csv（房屋价格数据，目标变量为 SalePrice）

环境：Python 3.x，Jupyter Notebook，依赖库：pandas、numpy、matplotlib、seaborn、scikit-learn

安装命令：pip install jupyter pandas numpy matplotlib seaborn scikit-learn

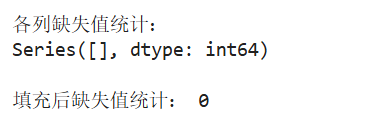
# 三、任务与思考

# （1）缺失值的检测与缺失值处理

思考：为何要区分数值型与类别型列进行填充？均值/众数/KNN/树模型填充的优缺点是什么？不同策略可能引入哪些偏差？

方法：数值型列以均值填充，类别型列以众数填充；可附加探索：使用 KNNImputer 对比效果。

关键代码：  
num\_cols = df.select\_dtypes(include=[np.number]).columns  
df.loc[:, num\_cols] = df[num\_cols].fillna(df[num\_cols].mean())  
cat\_cols = df.select\_dtypes(include=['object']).columns  
for c in cat\_cols:  
 df.loc[:, c] = df[c].fillna(df[c].mode()[0])



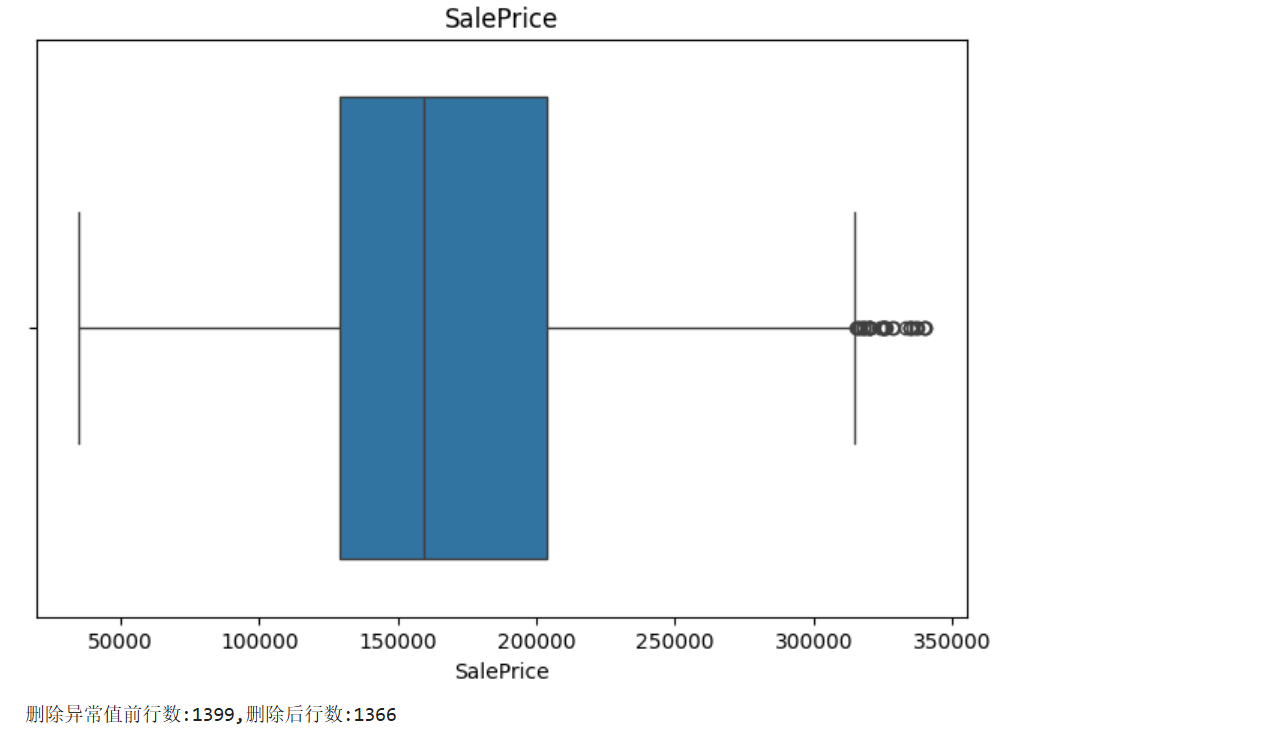
# （2）异常值检测

思考：为何先处理缺失值再处理异常值？为何以 SalePrice 为主进行异常值检测？IQR 与 Z-Score 的区别？

方法：以 IQR 方法为例：下界=Q1-1.5\*IQR，上界=Q3+1.5\*IQR；记录删除前后样本数。

关键代码：  
Q1, Q3 = df['SalePrice'].quantile(0.25), df['SalePrice'].quantile(0.75)  
IQR = Q3 - Q1  
before, after = df.shape[0], df[~((df['SalePrice'] < Q1-1.5\*IQR) | (df['SalePrice'] > Q3+1.5\*IQR))].shape[0]

SalePrice 箱线图/异常值可视化



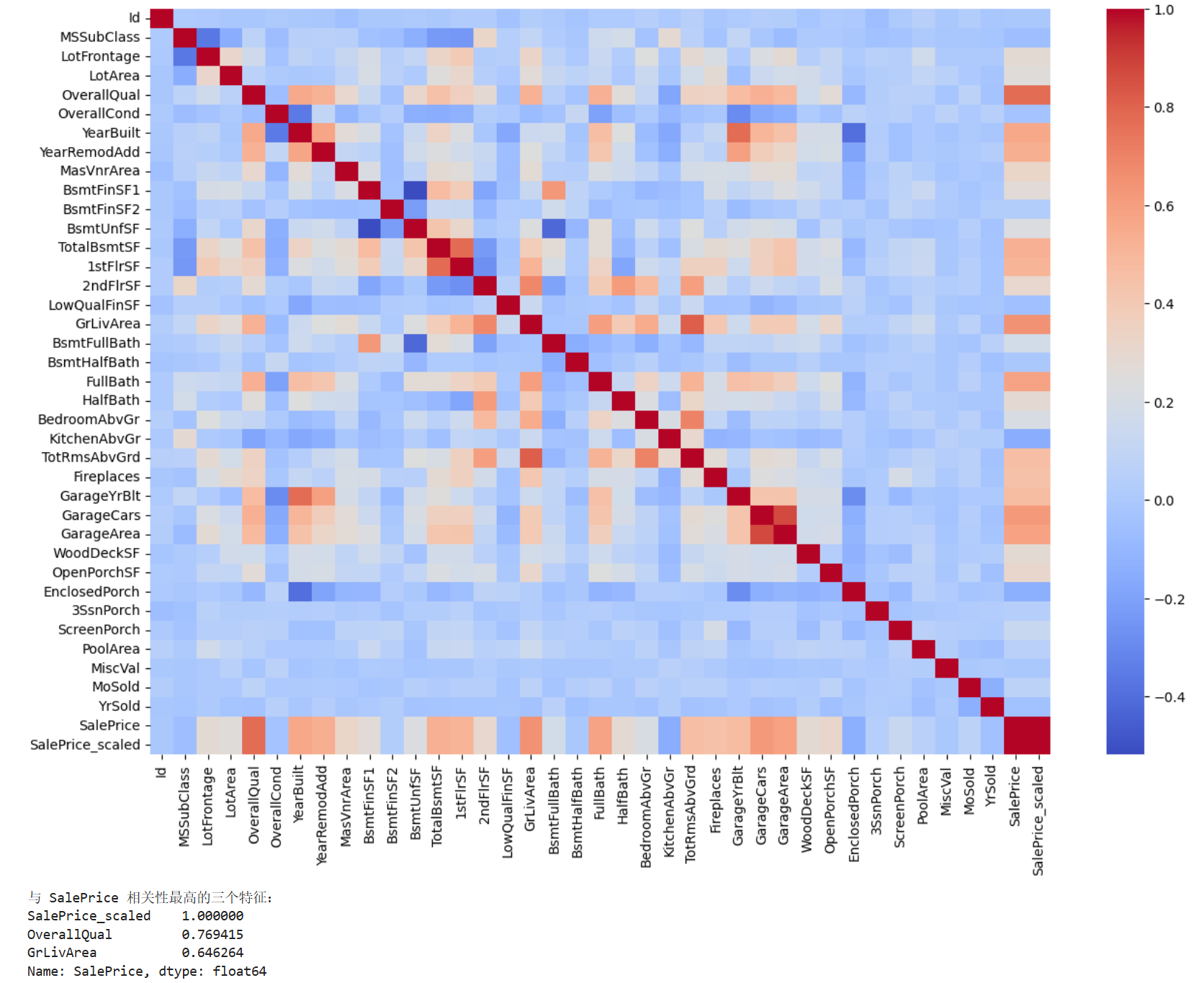
# （3）特征间的相关性分析

思考：为什么只用数值型列计算皮尔逊相关系数？类别变量该如何处理（独热编码/目标编码）？

方法：对数值型特征计算相关矩阵，绘制热力图；找出与 SalePrice 相关性最高的三个特征。

关键代码：  
num\_df = df.select\_dtypes(include=[np.number])  
corr = num\_df.corr()  
top3 = corr['SalePrice'].drop('SalePrice').abs().sort\_values(ascending=False).head(3)

相关性热力图



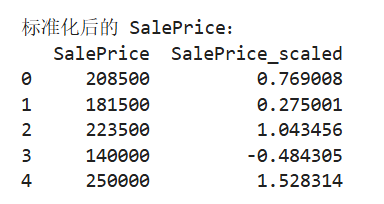
# （4）对 SalePrice 属性进行标准化

思考：为什么需要标准化？与归一化的区别？对后续建模（如正则化回归）的影响？

方法：使用 StandardScaler 对 SalePrice 标准化，注意使用 .loc 避免 SettingWithCopy 警告。

关键代码：  
from sklearn.preprocessing import StandardScaler  
scaler = StandardScaler()  
df.loc[:, 'SalePrice\_scaled'] = scaler.fit\_transform(df[['SalePrice']])

标准化前后对比表/图



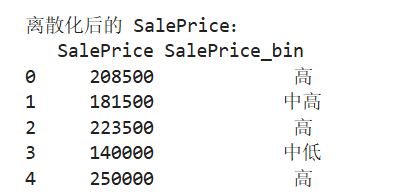
# （5）根据 SalePrice 属性进行离散化

思考：等频（qcut）与等宽（cut）离散化的区别？分箱对鲁棒性与解释性的影响？

方法：将 SalePrice 分为四档：低/中低/中高/高（等频分箱）。

关键代码：  
df.loc[:, 'SalePrice\_bin'] = pd.qcut(df['SalePrice'], q=4, labels=['低','中低','中高','高'])

分箱后频数直方图/箱线图

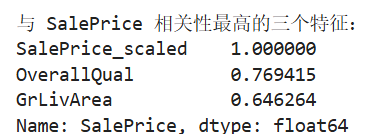


# （6）找出与 SalePrice 相关性最高的三个特征，并给出合理解释

分析结论：  
示例（以常见结果为例，实际请以你运行结果为准）：  
1) OverallQual：整体质量评分越高，通常房价越高（材料、做工、设计等综合影响）。  
2) GrLivArea：地上居住面积越大，房价通常越高（居住空间直接影响价值）。  
3) GarageCars/GarageArea：车库容量/面积越大，房价倾向更高（配套设施与便利性）。

注意事项：相关性不等于因果；可能存在多重共线性与非线性关系；可尝试对 SalePrice 取对数缓解偏态。

Top3 特征条形图/散点图



# 四、综合结论与反思

更深入与发散的思维：

- 结构性因素：位置/质量/面积的交互关系，可能存在的分层效应（如新区 vs 老区）。

- 统计性风险：异方差性、共线性、辛普森悖论，相关≠因果。

- 方法改进：对数变换、类别编码（One-Hot）、交互项、KNN/树模型填充、稳健缩尾（Winsorization）。

# 五、附录A：运行与复现实验

1）安装：pip install jupyter pandas numpy matplotlib seaborn scikit-learn

2）运行：在第二次作业目录执行 jupyter notebook，打开 lab2.ipynb 依次运行所有单元。

3）版本：Python 3.x，库版本请以你本地环境为准并在此处备注。