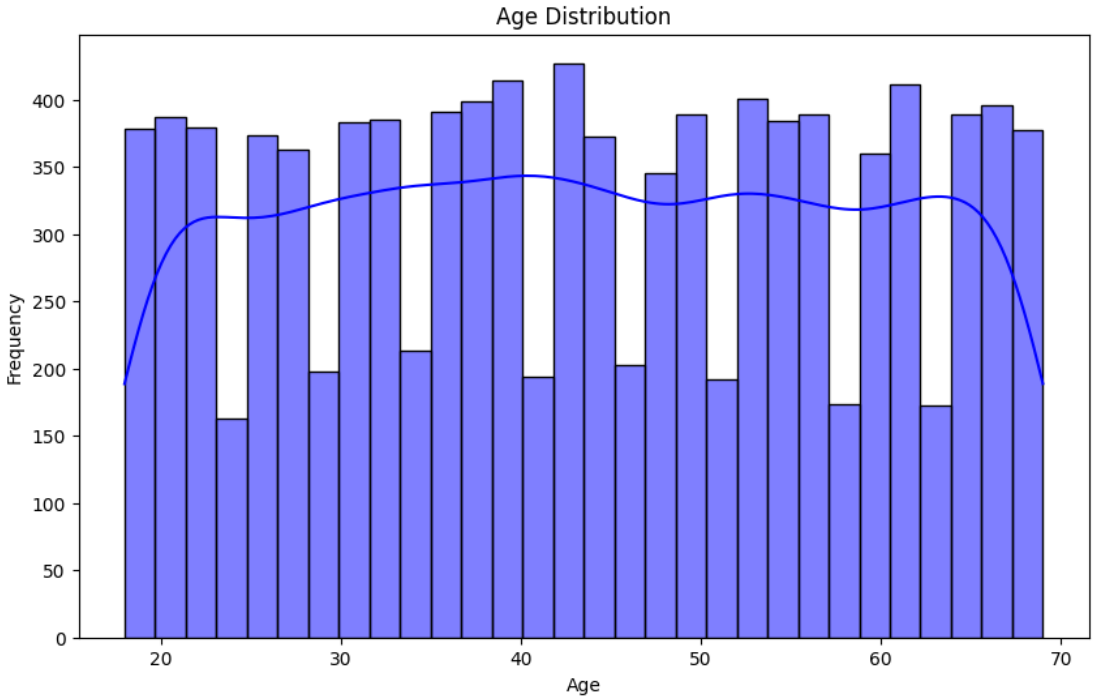


一、年龄分布直方图（Age Distribution）

1、图表内容

横轴：年龄（Age），范围为 20-70 岁，以 10 岁为间隔划分区间。

纵轴：频率（Frequency），表示每个年龄区间内的样本数量。

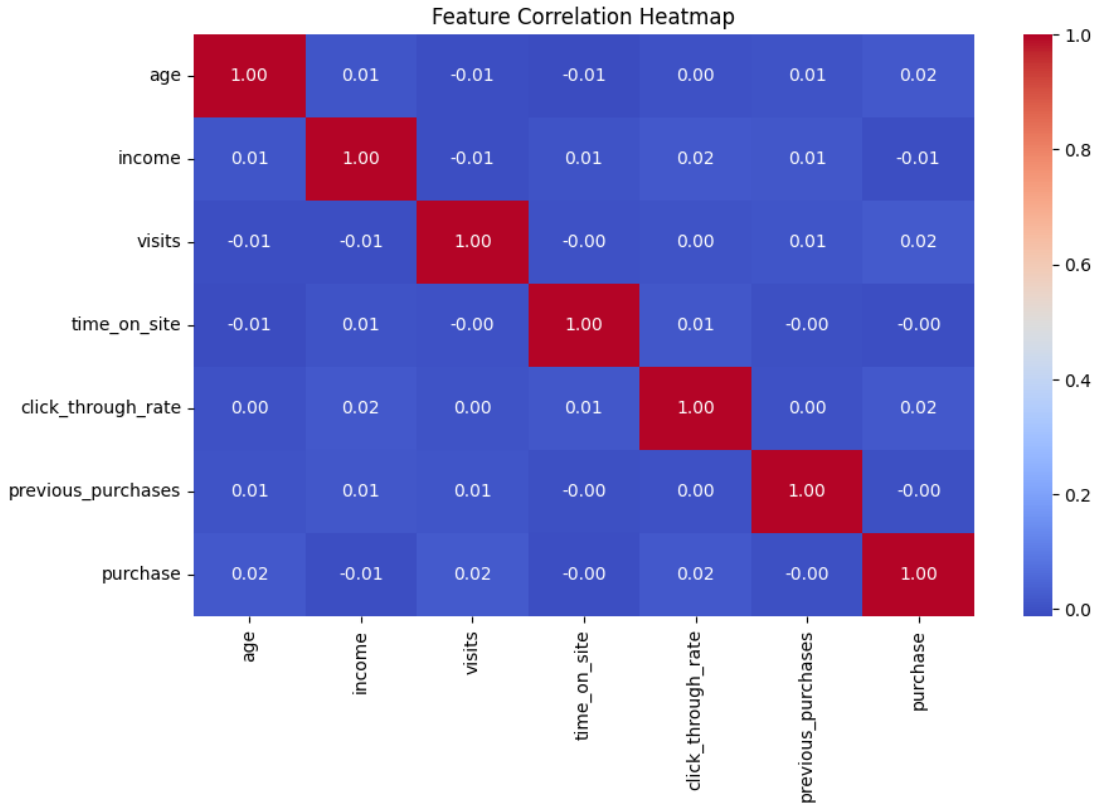
曲线：核密度估计曲线（KDE），平滑展示年龄分布的概率密度。

2、分析

数据生成逻辑：代码中通过 np.random.randint(18, 70, size=data_size)生成 18-70 岁的随机整数，因此理论上年龄分布应为均匀分布。

图表表现：直方图呈现近似均匀分布，各年龄区间的频率差异较小，KDE 曲线平滑且无明显峰值，符合随机生成的预期。

作用：直观展示用户年龄的分布范围和集中趋势，帮助理解数据的离散程度，为后续特征工程提供基础信息。



二、特征相关性热图（Feature Correlation Heatmap）

1、图表内容

矩阵元素：各特征之间的皮尔逊相关系数（数值范围 [-1, 1]），正数表示正相关，负数表示负相关。

颜色映射：使用 coolwarm 色系，红色表示正相关，蓝色表示负相关，颜色越深绝对值越大。

2、关键数值（根据代码注释推断）：

年龄（age）与其他特征的相关系数接近 0（如与收入 income 的相关系数为 0.01），表明年龄与其他特征线性相关性极弱。各特征之间的相关系数普遍接近 0（绝对值均小于 0.02），说明特征之间相互独立，无显著线性关联。

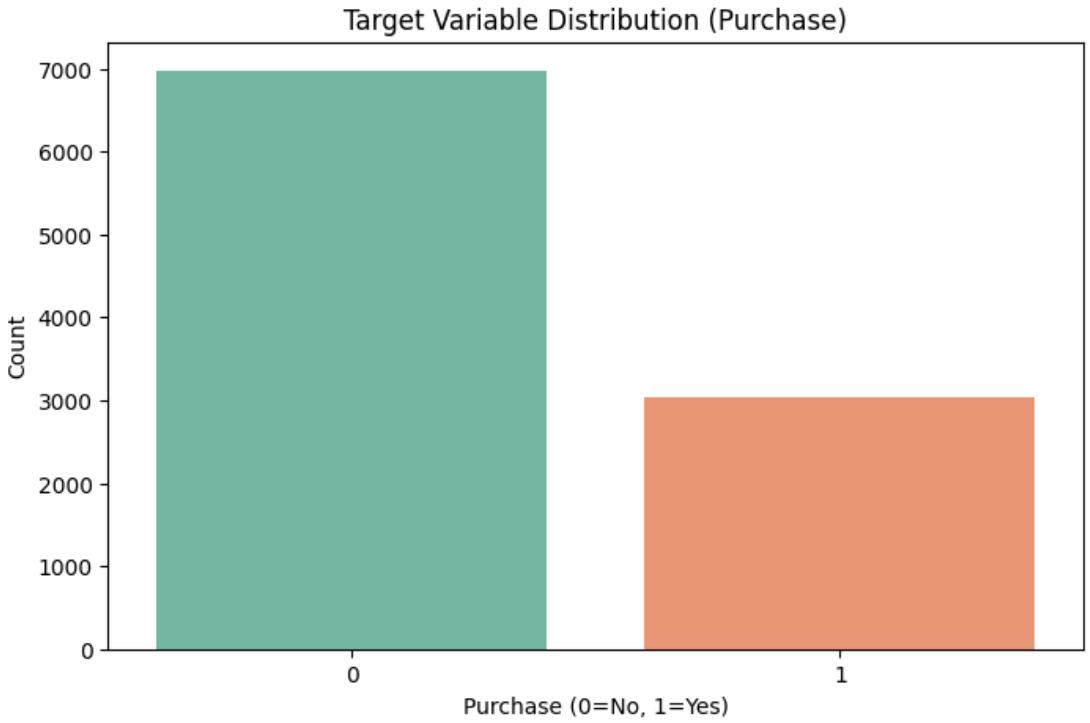
3、分析

数据生成逻辑：代码中各特征（如收入、访问次数等）均为独立随机生成（如泊松分布、均匀分布等），因此理论上特征间无实际相关性。

图表表现：热图中所有单元格数值接近 0，颜色趋近于白色，验证了特征独立性。

作用：帮助识别特征间的多重共线性问题，本例中无强相关特征，无需进

行特征筛选或降维。



三、目标变量分布（Target Variable Distribution (Purchase)）

1、图表内容

横轴：购买行为（Purchase），0 表示未购买，1 表示购买。纵轴：数量（Count），表示各类别的样本数。

2、数据表现：

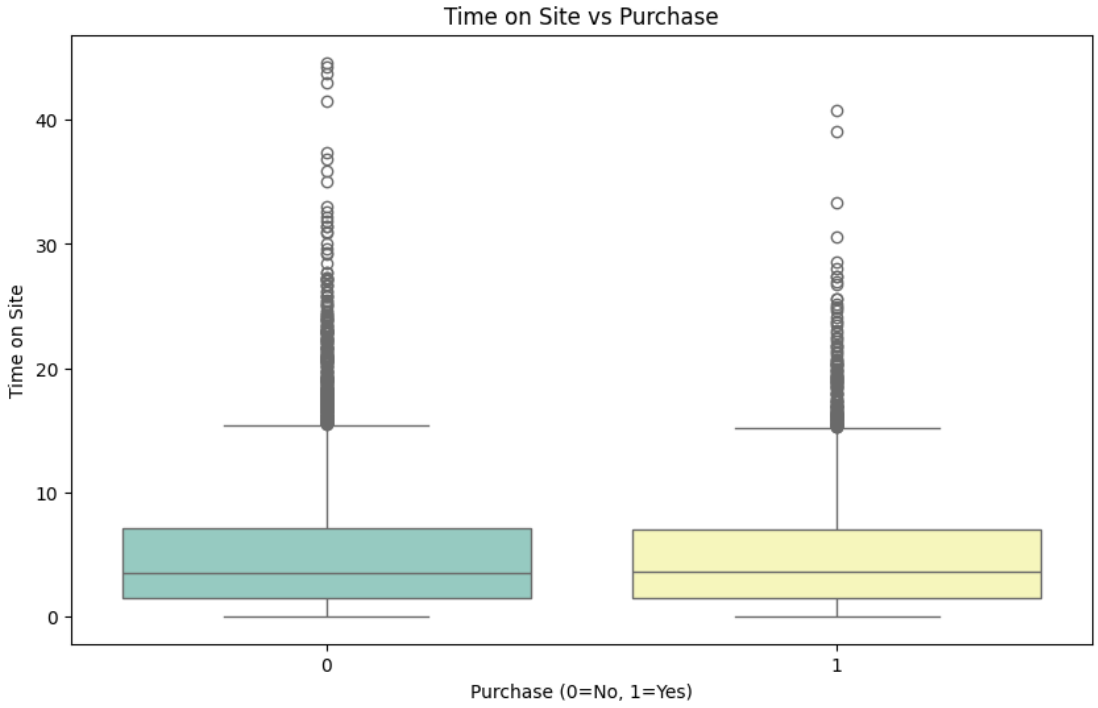
未购买（0）的样本数约为 7000，购买（1）的样本数约为 3000，比例为 7:3。

3、分析

数据生成逻辑：代码中通过 np.random.choice([0, 1], p=[0.7, 0.3])生成购买行为，明确设置购买概率为 30%，因此图表结果与生成逻辑一致。

关键问题：目标变量存在轻度类别不平衡（未购买样本占多数），可能影响分类模型的性能（如倾向于预测负类），需在模型评估中使用适合不平衡数据的指标（如 AUC-ROC）。

作用：快速验证数据生成的合理性，识别类别不平衡问题，为模型选择（如使用加权损失函数）提供依据。



四、停留时间与购买关系（Time on Site vs Purchase）

1、图表内容

横轴：购买行为（Purchase），0 和 1 分别表示未购买和购买。纵轴：网站停留时间（Time on Site），单位为分钟（假设）。

2、箱线图元素：

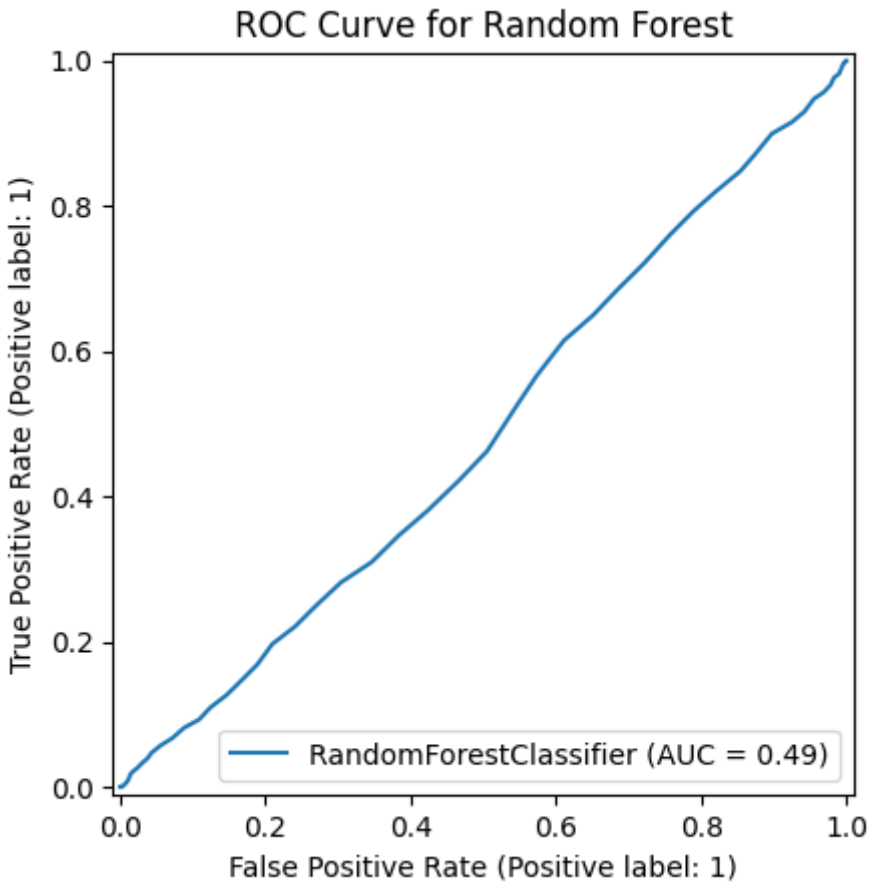
中位数：购买组（1）的停留时间中位数略高于未购买组（0）。四分位距（IQR）：两组的 IQR 相近，数据分散程度相似。异常值：两组均存在少量停留时间极长的异常值。

3、分析

数据生成逻辑：停留时间由指数分布 np.random.exponential(5)生成，理论上均值为 5，且指数分布具有长尾特性（少数样本值极大），与图表中的异常值现象一致。

潜在关系：虽然购买组的中位数略高，但两组的分布重叠较多，说明停留时间与购买行为的相关性较弱（与热图中相关系数接近 0 的结论一致）。

作用：辅助判断特征与目标变量的关联性，本例中停留时间对购买行为的预测价值有限，可能不是关键特征。



五、随机森林 ROC 曲线（ROC Curve for Random Forest）

1、图表内容

横轴：假正率（False Positive Rate, FPR），表示误将负类预测为正类的比例。纵轴：真正率（True Positive Rate, TPR），表示正确识别正类的比例。曲线：随机森林模型的 ROC 曲线，AUC 值为 0.49（接近 0.5）。

2、分析

模型性能：AUC=0.49 接近 0.5，表明模型预测效果仅略优于随机猜测（随机猜测的 AUC 为 0.5），几乎无实际预测价值。

3、原因分析：

数据生成时特征与目标变量（购买行为）无实际关联（特征均为随机生成，与 purchase 列无因果关系），导致模型无法学习到有效模式。

类别不平衡问题可能加剧了模型对负类的偏向，但根本原因是特征无信息量。

作用：通过可视化评估模型在不同阈值下的分类能力，本例中曲线接近对角线，验证了模型的无效性。

总结：图表与代码的关联

数据生成的随机性：所有图表均基于随机生成的数据，特征与目标变量无真实关联，因此：特征相关性热图显示低相关性。模型性能（AUC=0.49）接近随机水平。可视化目的：代码通过图表演示数据探索和模型评估的基本流程，而非展示真实业务规律。实践意义：若在真实场景中出现类似结果，需检查数据质量、特征工程是否有效，或重新选择模型。