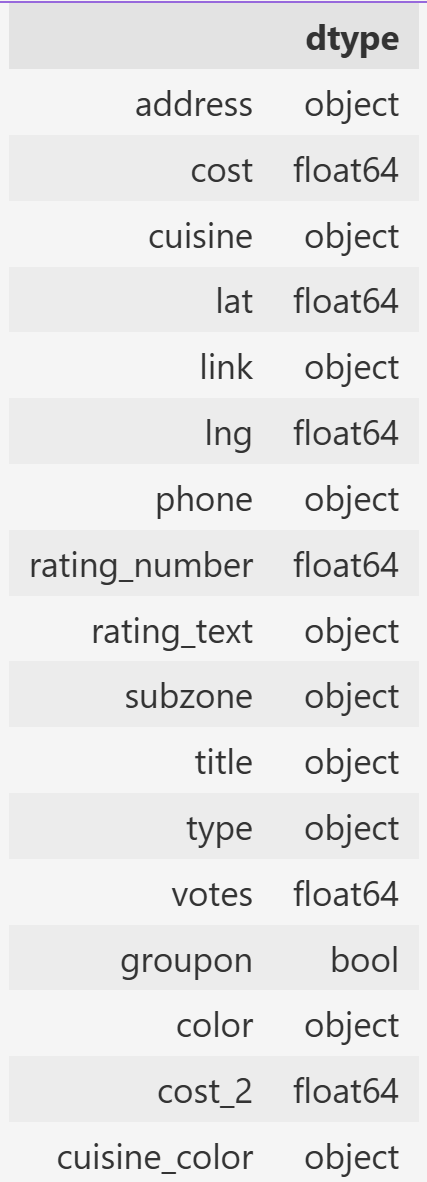
Assignment1

PartA

**EDA 主要图表**：独特菜系、Top3 郊区、Excellent vs Poor 成本差异、相关性、1 个交互图说明优势。







Initial Data Exploration

The dataset contains 10,500 rows and 17 columns. Missing values are concentrated in the fields rating\_number, votes, and rating\_text(about 31.6% missing), while cost and cost\_2 have around 3.3% missing, and lat/lng about 1.8%. Other fields are almost complete.

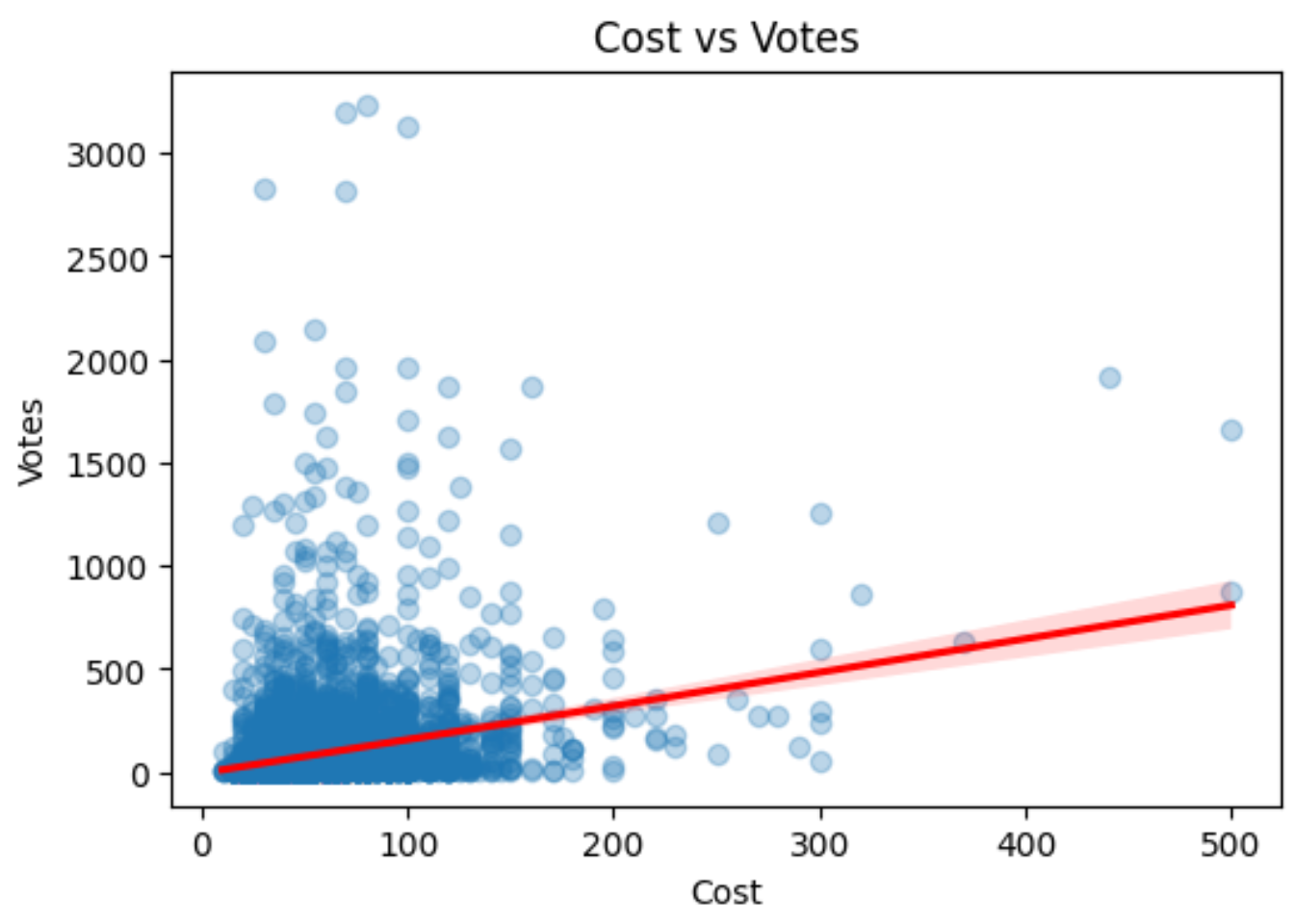
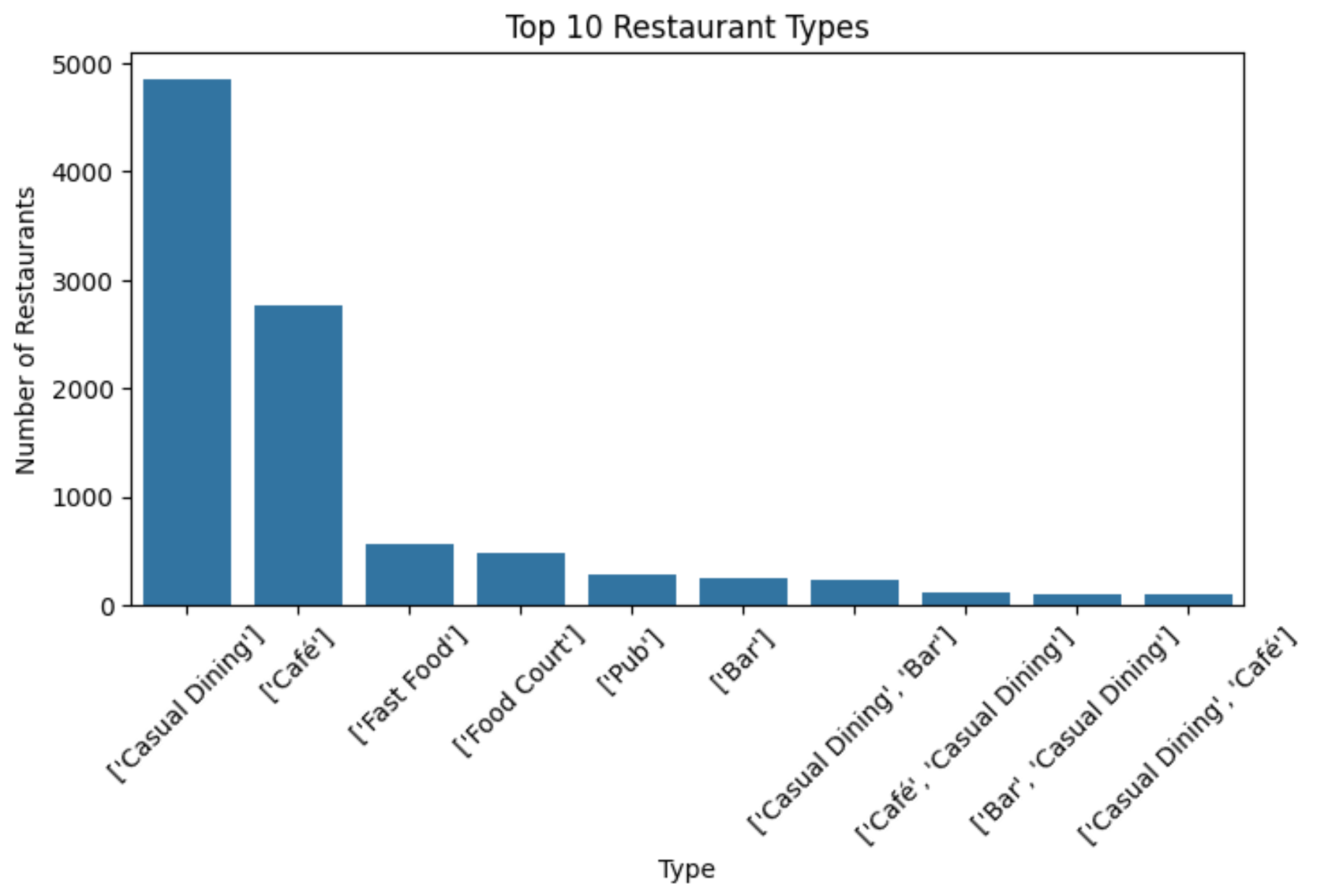
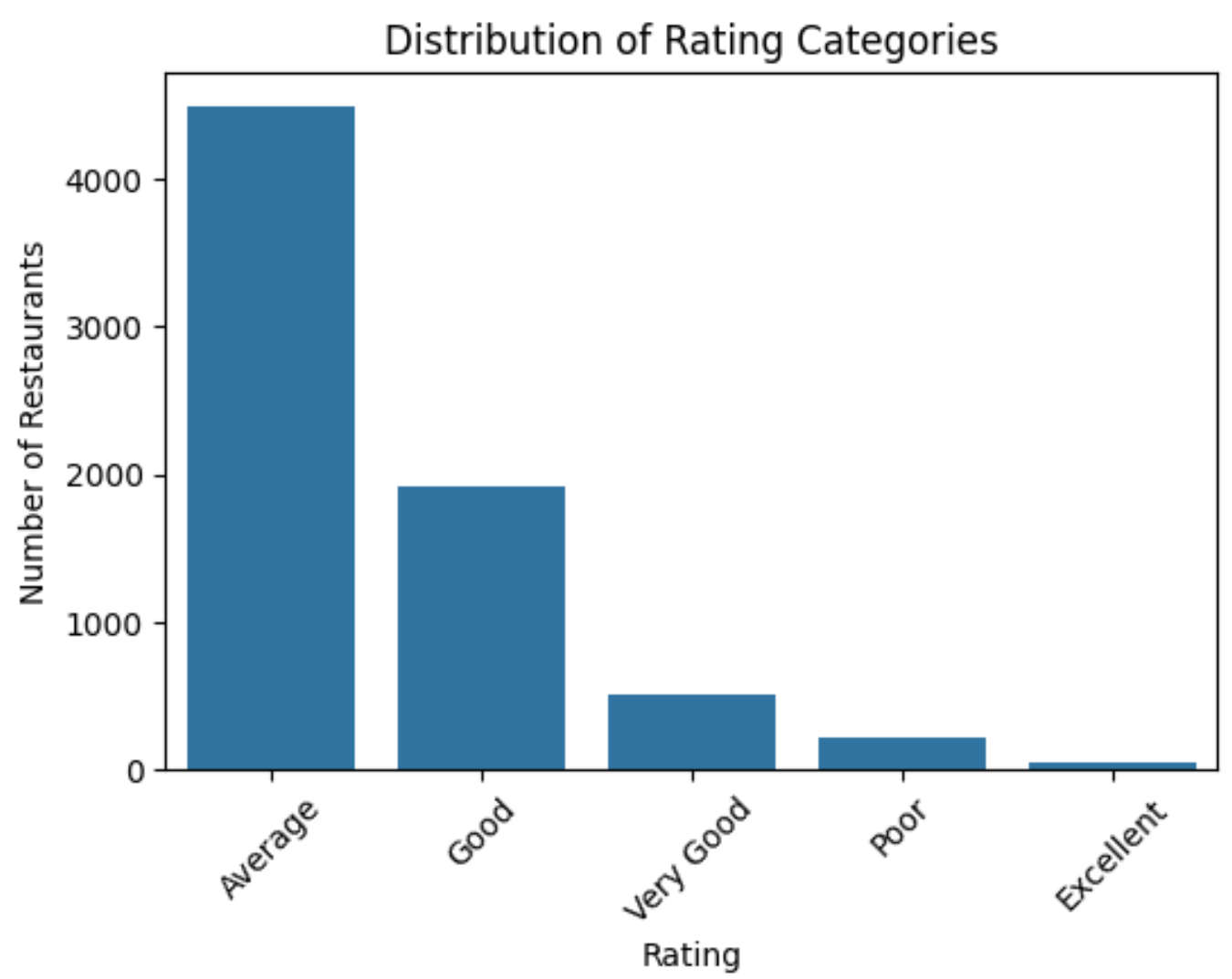
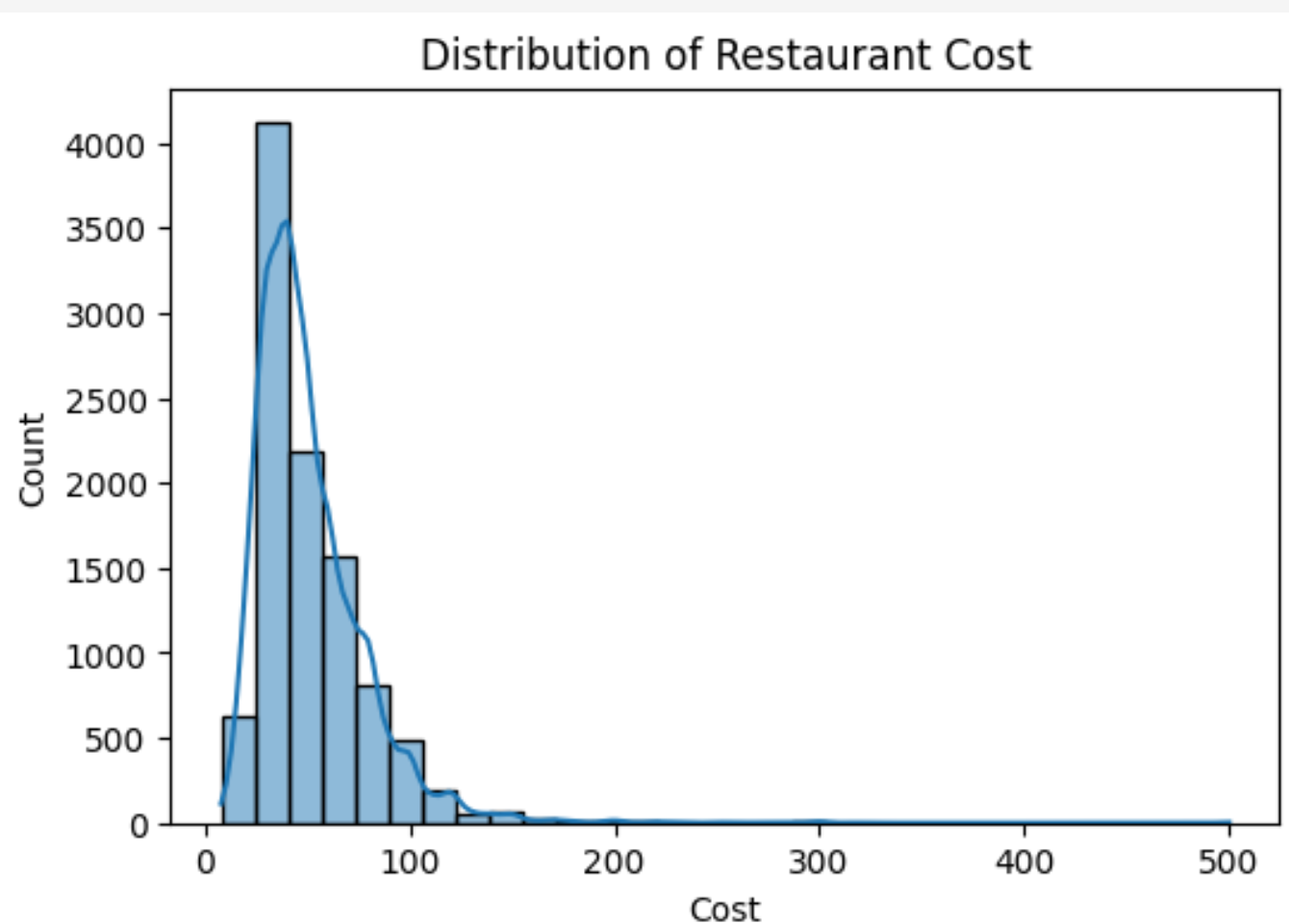
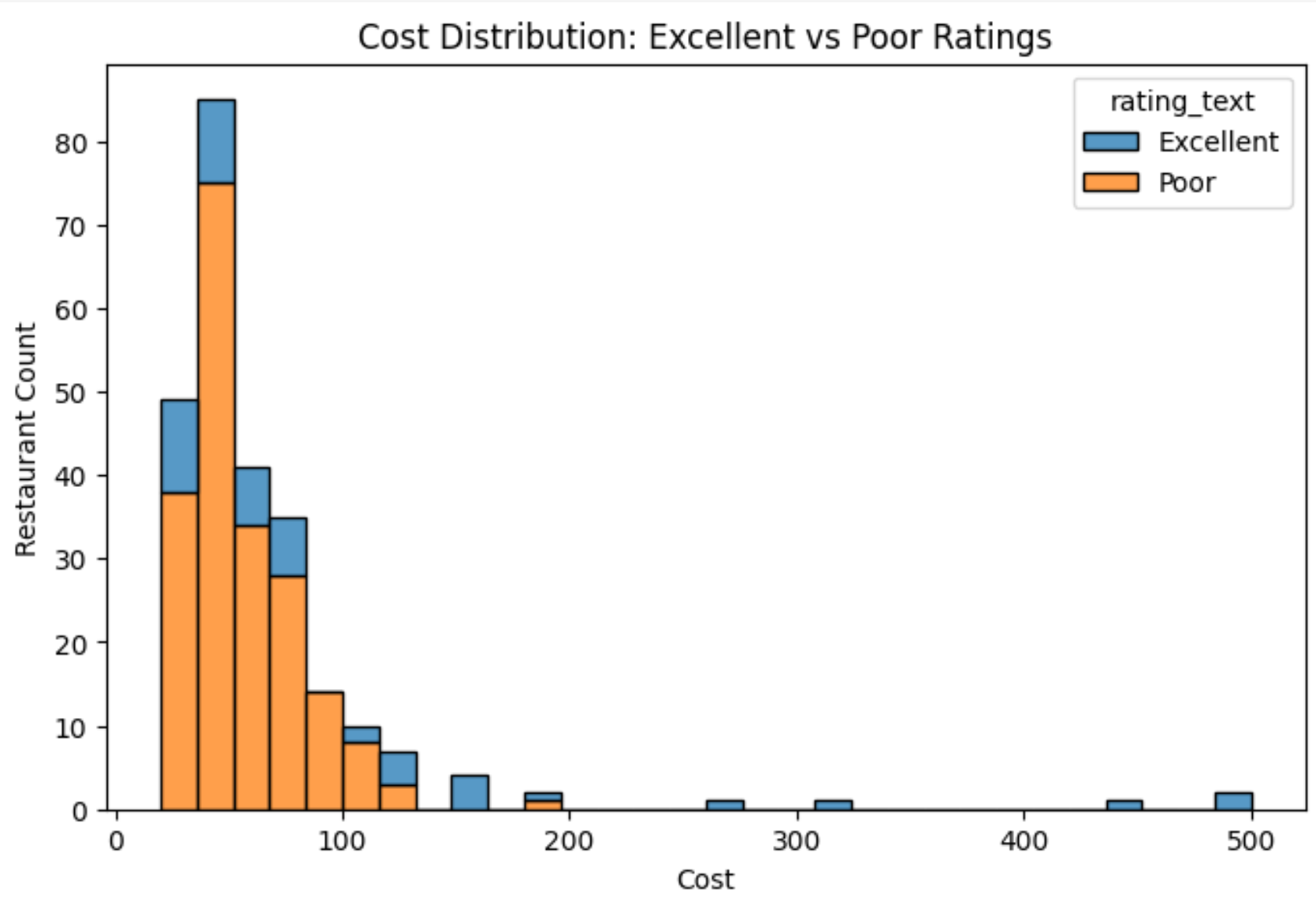
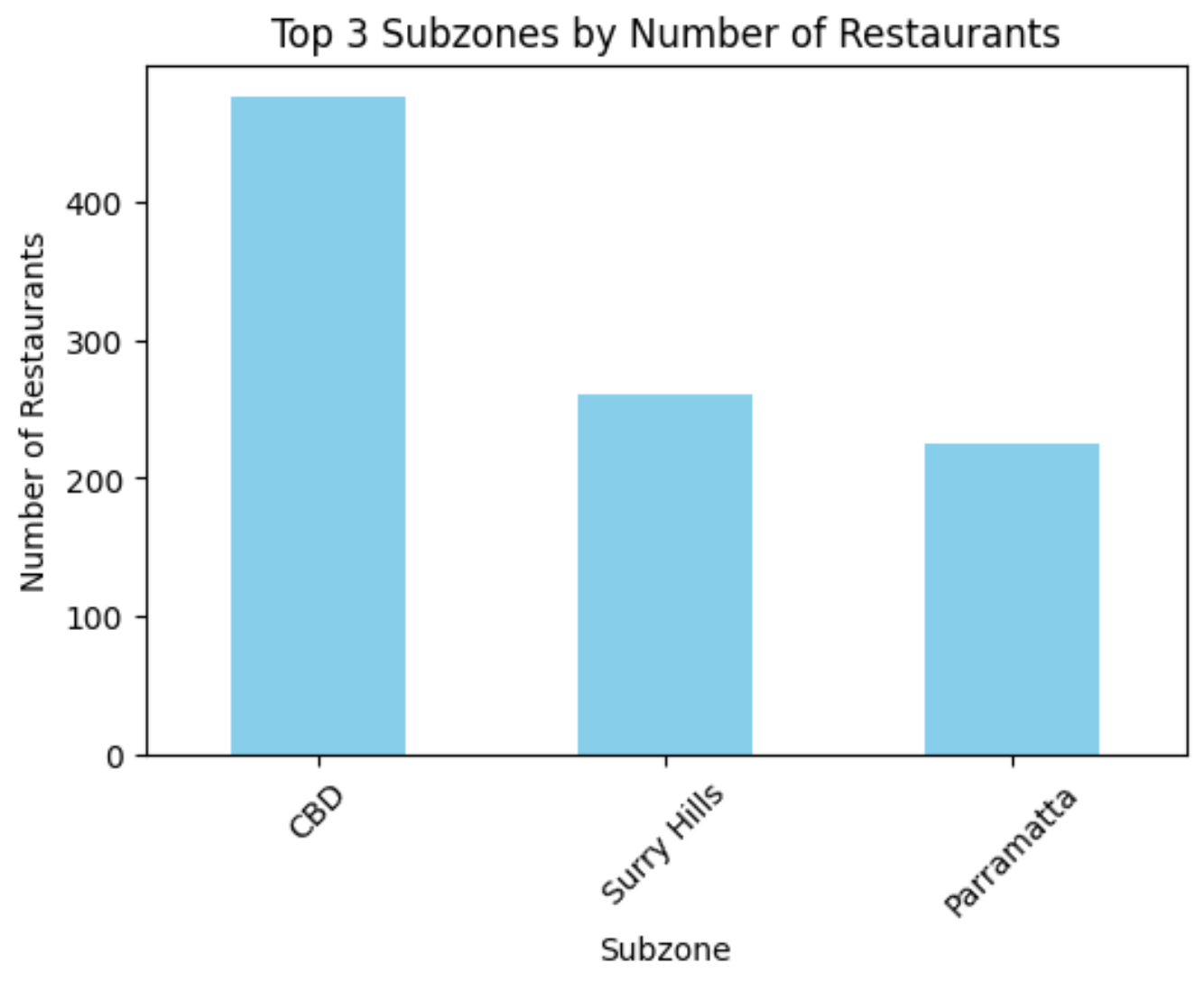
Numeric features show clear variation:

votes is highly skewed with a long tail (mean is about 84, median is about 32, maximum = 5000).

cost values range from 8 to 800, indicating potential outliers.

rating\_number is centered around 3.3 with relatively low variance.

Categorical features demonstrate high cardinality. For example, cuisine includes 1,759 unique values,subzone has 572, and type has 66 categories. This suggests that feature engineering will be necessary for modeling.

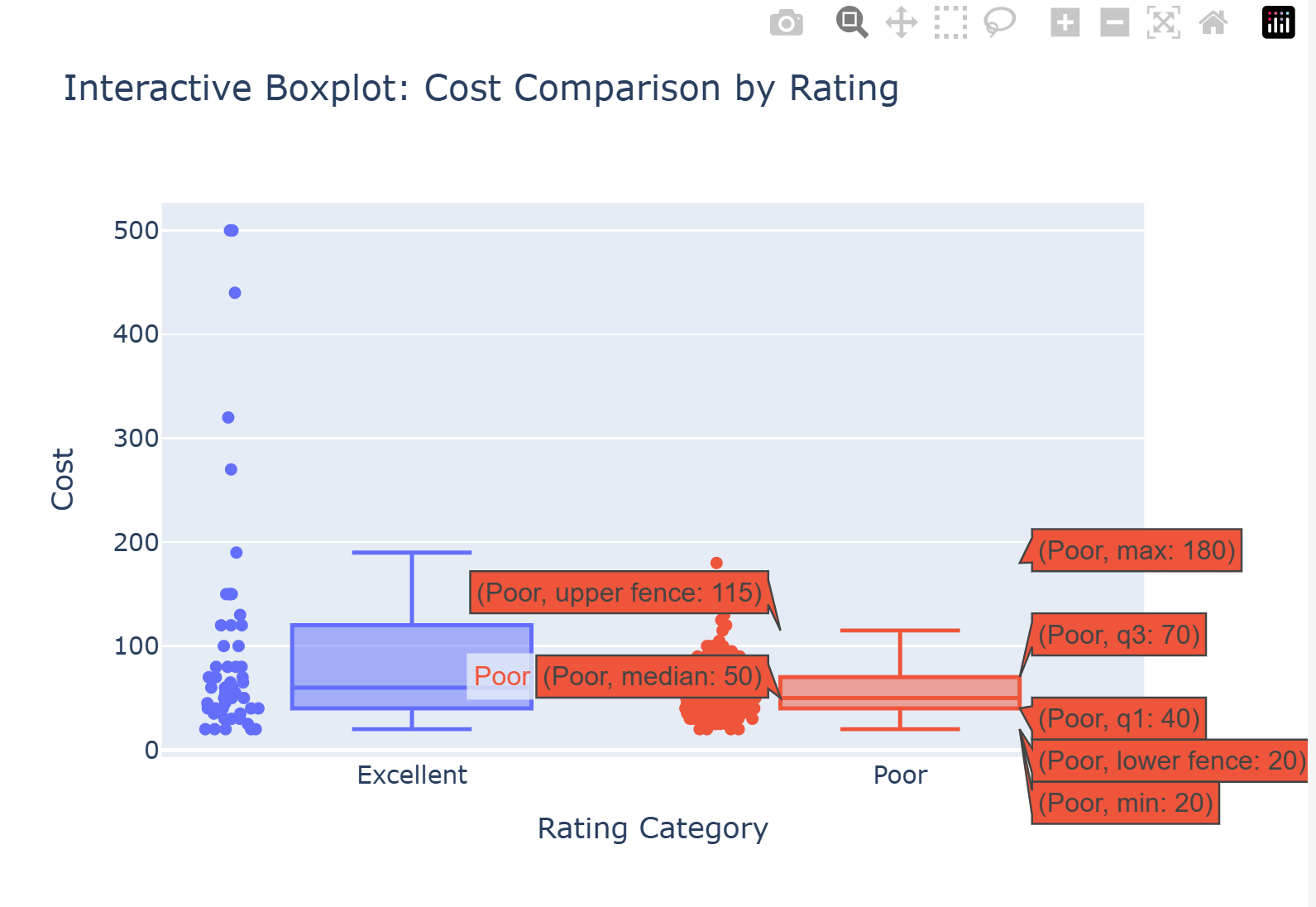
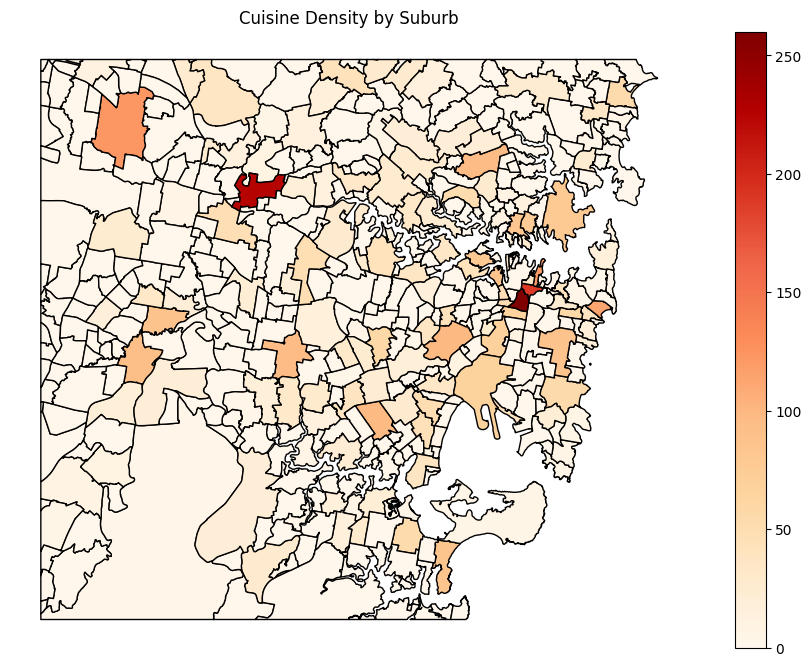
Key Variable Exploration and Observations:

The distribution of restaurant cost reveals a strong right skew, indicating that while most restaurants have moderate pricing (centered around $45), there are a few high-end outliers charging up to $800. This long-tail behavior suggests price-sensitive clustering in the mid-to-low cost range.

Rating distributions show a disproportionate number of restaurants receiving high ratings (especially “Excellent”), suggesting either genuine service quality or potential rating inflation. In contrast, lower-rated categories are much less represented.

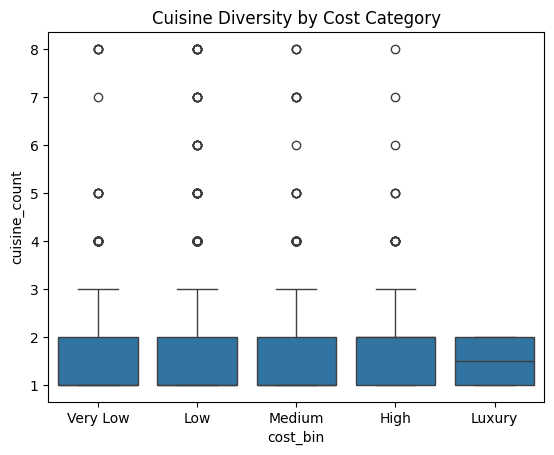
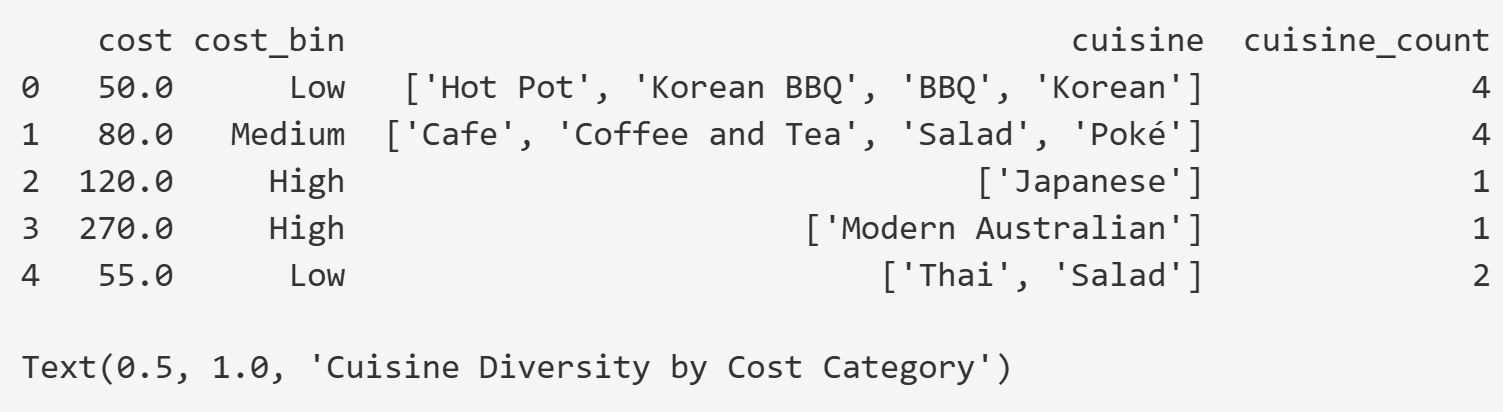
When exploring restaurant types, Casual Dining and Café dominate the dataset, far outnumbering other formats such as Fast Food, Bar, or Food Court. This indicates a consumer or business preference toward these formats in the observed geography.

The relationship between cost and votes displays a weak but positive correlation. Higher-cost restaurants tend to receive more votes, possibly reflecting greater visibility, larger customer bases, or increased marketing. However, the presence of numerous low-cost restaurants with high vote counts suggests affordability does not necessarily equate to low popularity.



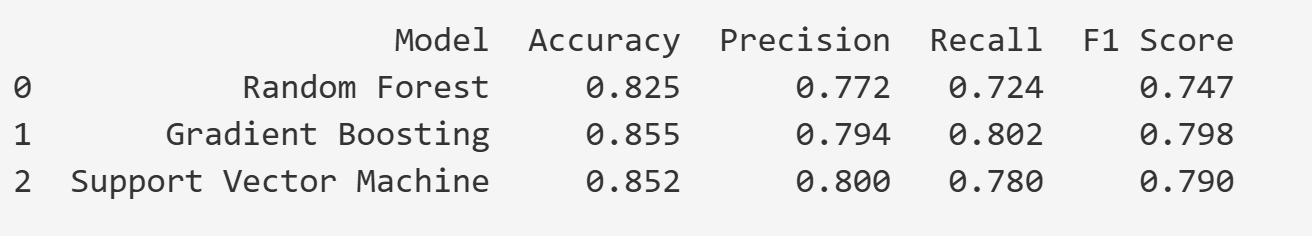
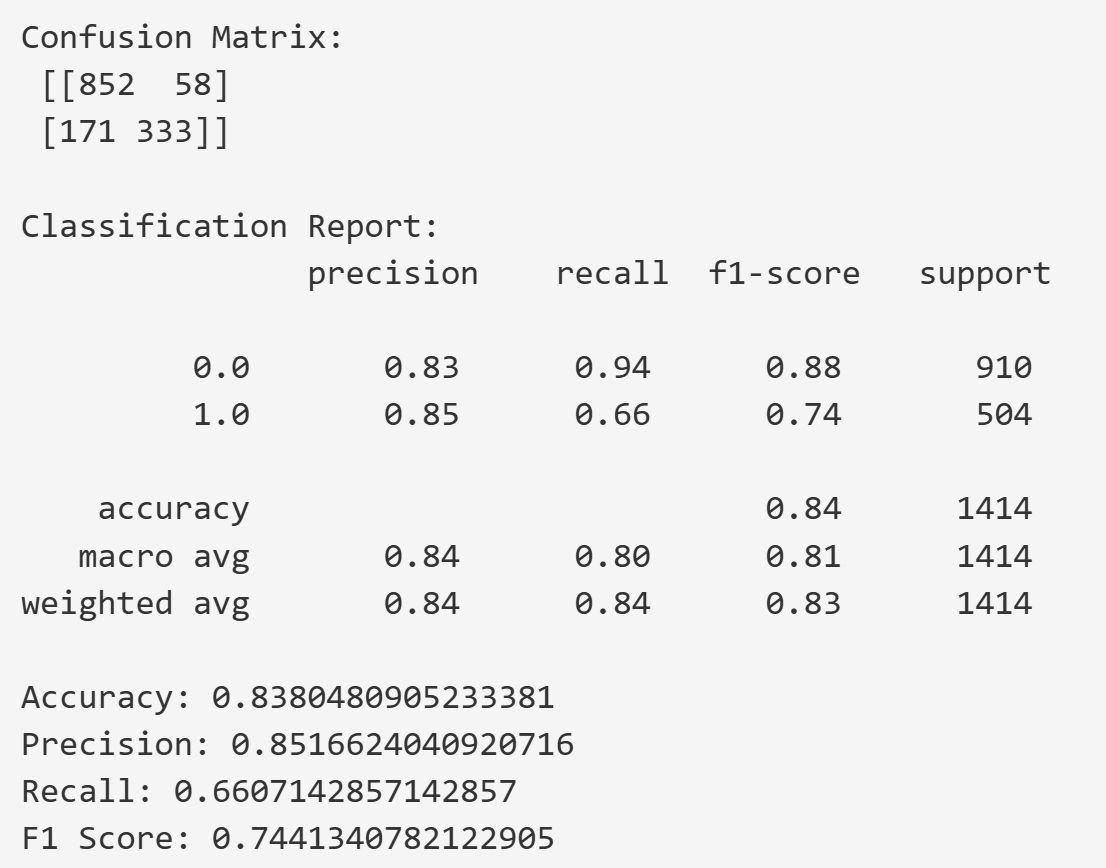
Interactive visualizations significantly enhance data exploration by allowing users to zoom, hover, and identify outliers dynamically. Compared to the static boxplot, the Plotly version allows for detailed inspection of cost distributions across rating categories, making it ideal for both analytical presentations and business dashboards.

PartB

Model A (LinearRegression) MSE: 0.13430041386272504Model

B (Gradient Descent) MSE: 2.4268453428824253

We compared two models for predicting restaurant ratings: Model A, which uses Scikit-learn's built-in LinearRegression, and Model B, a custom implementation based on gradient descent. The Mean Squared Error (MSE) of Model A was 0.1343, while Model B had a significantly higher MSE of 2.4268. This result clearly indicates that Model A performed much better in terms of predictive accuracy. The poor performance of Model B may be due to factors such as an inappropriate learning rate, insufficient number of iterations, or lack of proper feature scaling. Overall, Model A is more reliable and efficient for this task, and its low error demonstrates the advantage of using optimized, library-based regression solutions over manually implemented gradient descent methods.

In this binary classification task, the Gradient Boosted trees model performed best overall, leading in precision, recall, and F1 score, making it suitable for subsequent deployment or further optimization.

In the regression task, the Scikit-Learn Linear Regression achieved a slightly lower Mean Squared Error (MSE = 0.1343) compared to PySpark Linear Regression (MSE = 0.1351), indicating that Scikit-Learn performs marginally better on small to medium-sized datasets. For the classification task, Scikit-Learn Logistic Regression reached an accuracy of 0.8380, while PySpark Logistic Regression achieved 0.8340. The difference is minimal, but Scikit-Learn additionally provides detailed metrics such as precision, recall, and F1-score, making it more flexible and informative for model evaluation on smaller datasets. Regarding speed, Scikit-Learn completed the training almost instantly (0.0s–0.1s), whereas PySpark required approximately 37s due to the overhead of initializing the distributed computing framework. However, in terms of scalability, PySpark has a significant advantage, as it is designed to handle large-scale data efficiently across distributed environments. Therefore, Scikit-Learn is more suitable for local experiments and small to medium datasets, while PySpark is preferable when working with big data and distributed computing scenarios.

PartC

# 第 12 步：PDF 报告怎么写（结构建议）



**回归对比表**（MSE：sklearn LR vs 你实现的 GD）

**分类对比表**（LogReg + 3 模型：Precision/Recall/F1/Acc）

**PySpark vs Scikit-Learn 的反思**（准确率、易用性、规模化/速度）

**可复现与协作**：列 Git/LFS/DVC 核心命令与 dvc.yaml 四阶段说明。

**仓库链接**（要求必须包含全部代码与 DVC 设置）。