

**实验(实训)课程报告**

**课程 数据挖掘技术实践**

|  |  |  |  |
| --- | --- | --- | --- |
| 院　　别 | 信息技术与管理 | 专 业 | 计算机科学与技术 |
| 届 别 | 2022 | 班 级 | 计算机科学与技术（职高）一班 |
| 学生姓名 | 邱宇杰 | 学 号 | 202205650121 |
| 指导教师 | 钟敏娟 | 职 称 |  |

湖南财政经济学院

2024年 8 月

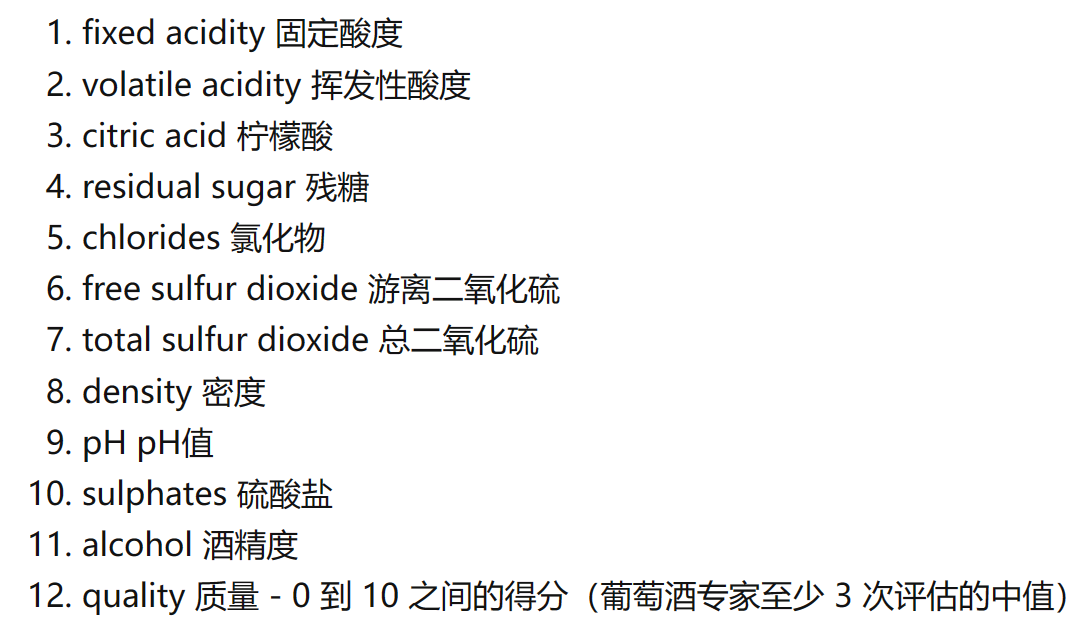
**实验(实训)报告1**

实验(实训)时间: 2024 年 10 月 18 日 指导教师评分：

|  |  |
| --- | --- |
| 实验项目 | 实验一：认识数据—红酒品质数据的可视化 |
| 组员名单 | 独立完成或姓名（学号）（操作性和验证性实验均独立完成，综合性实验1-2人一组） |
| 实验类型 | ■基础性实验 □验证性实验 □综合性实验 |
| 实验(实训)时间 | 2024年 月 日 |

1. 实验目的和要求

今有来自UCI机器学习实验室的葡萄酒质量数据集，分别提供了有关葡萄牙“青酒”的红葡萄酒和白葡萄酒变体的样本信息，每种样本都由专家做了质量评级，并进行了理化指标检验。包含以下12个字段：



要求以红葡萄酒(winequality-red.csv)为对象，对其中数据进行如下初步处理：

(1) 以分号为分隔符进行数据的读入。

(2) 初识数据，给出数据的基本信息，包括字段名称、类型、字段个数、字段是否缺失以及数值数据的统计参数值(均值、个数、最大、最小、标准差等)。

(3) 利用subplot绘制所有字段变量的箱线图、直方图、累计曲线图；绘制变量百分比饼图；分别绘制带注释和对角线矩阵的热力图，并观察各个字段的分布特点。

2．实验步骤（记录程序代码、分析实验结果）

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

# 读取数据，以分号为分隔符

file\_path = '作业\winequality-red.csv'

wine\_data = pd.read\_csv(file\_path, delimiter=';')

# 1. 显示数据集的基本信息和前5行数据

print("数据集的基本信息：")

print(wine\_data.info())

print("\n数据集的前5行：")

print(wine\_data.head())

# 2. 数值型字段的统计信息

print("\n数值型字段的统计信息：")

print(wine\_data.describe())

# 3. 数据可视化

# (1) 绘制各字段的箱线图

plt.figure(figsize=(20, 15))  # 设置图形大小

for i, column in enumerate(wine\_data.columns[:-1]):

    plt.subplot(3, 4, i + 1)  # 创建子图

    plt.boxplot(wine\_data[column].dropna())  # 绘制箱线图

    plt.title(f"{column} Boxplot")  # 设置标题

plt.tight\_layout()  # 调整子图间距

plt.show()  # 显示图形

# (2) 绘制直方图

plt.figure(figsize=(20, 15))  # 设置图形大小

for i, column in enumerate(wine\_data.columns[:-1]):

    plt.subplot(3, 4, i + 1)  # 创建子图

    plt.hist(wine\_data[column].dropna(), bins=15)  # 绘制直方图

    plt.title(f"{column} Histogram")  # 设置标题

plt.tight\_layout()  # 调整子图间距

plt.show()  # 显示图形

# (3) 绘制累计曲线图

fig, ax = plt.subplots(figsize=(10, 6))

for column in wine\_data.columns[:-1]:

    ax.plot(np.cumsum(wine\_data[column].dropna()), label=column)  # 绘制累计曲线

ax.set\_title('Cumulative Sum of Features')  # 设置标题

ax.legend(loc='upper left')  # 显示图例

plt.show()  # 显示图形

# (4) 绘制字段相关性的热力图

corr\_matrix = wine\_data.corr()

plt.figure(figsize=(12, 8))

plt.imshow(corr\_matrix, cmap='coolwarm', interpolation='none')

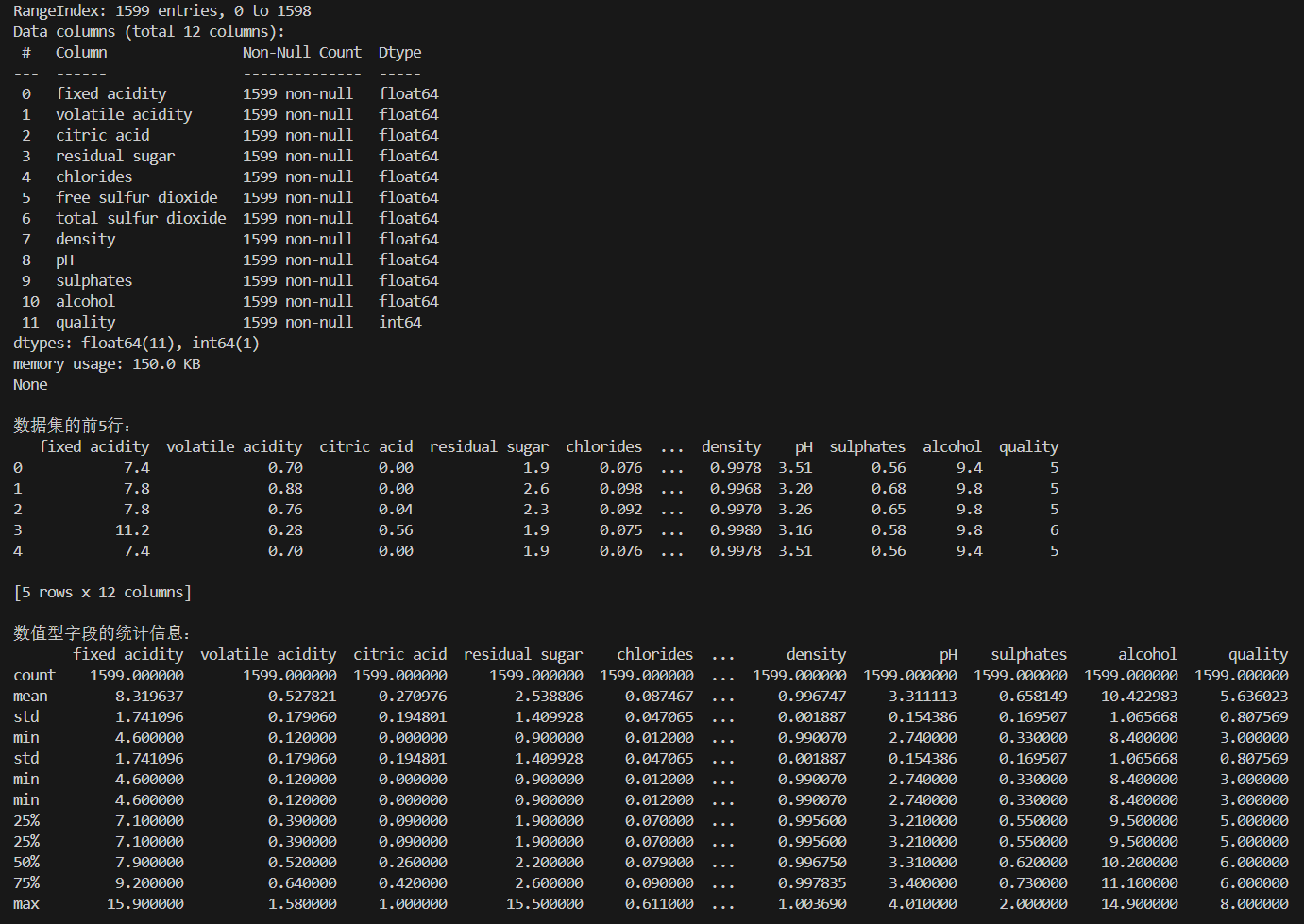
plt.colorbar()

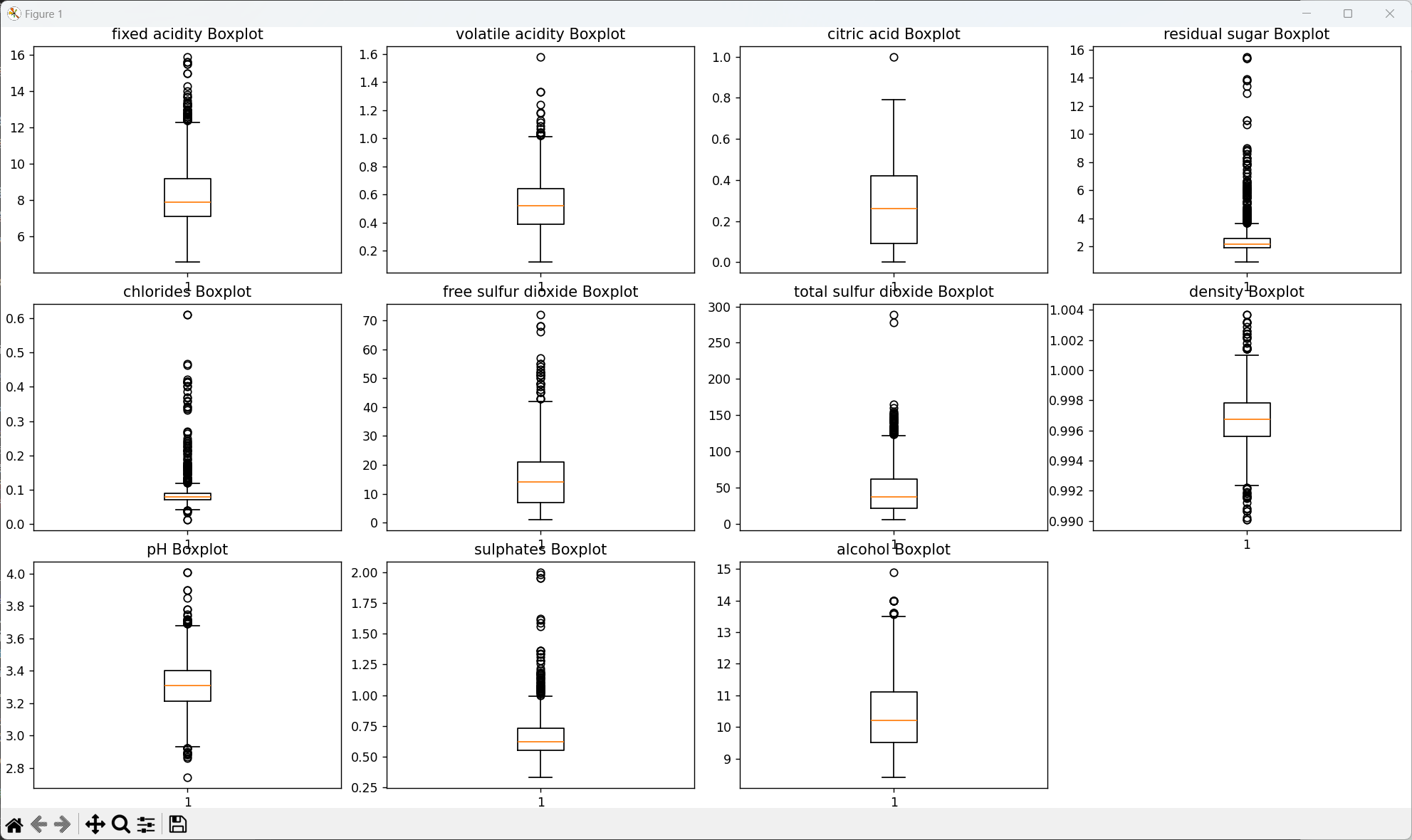
plt.xticks(np.arange(len(corr\_matrix)), corr\_matrix.columns, rotation=90)

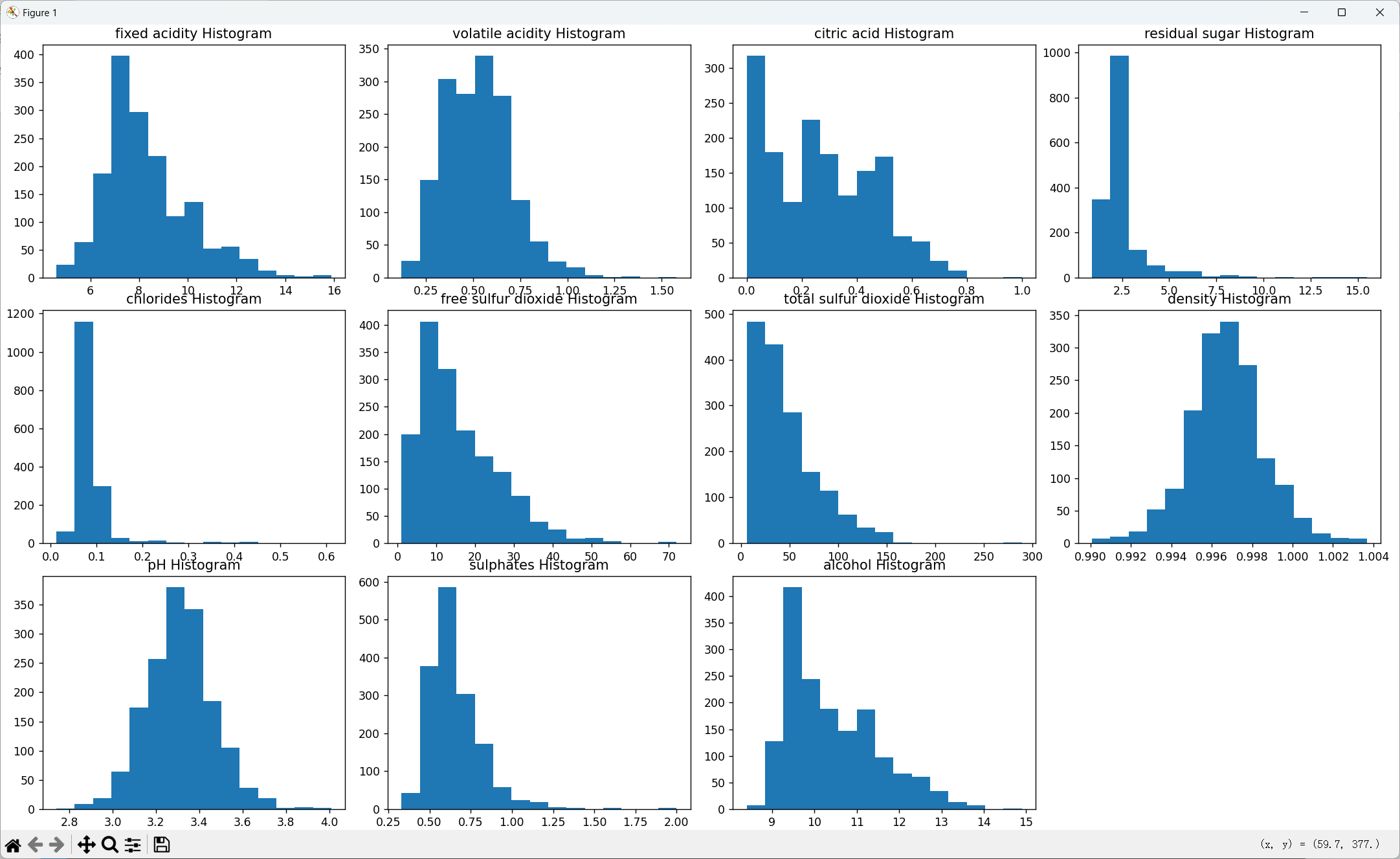
plt.yticks(np.arange(len(corr\_matrix)), corr\_matrix.columns)

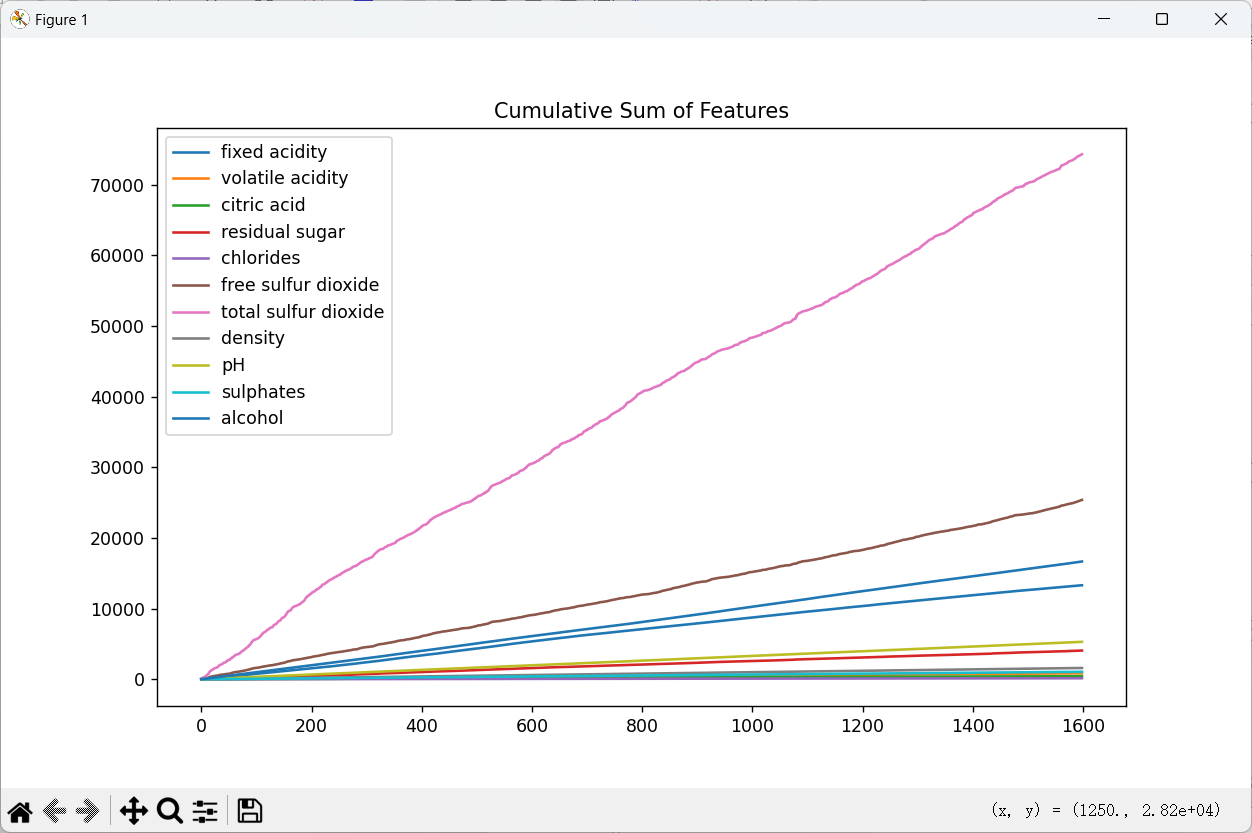
plt.title('Correlation Heatmap')

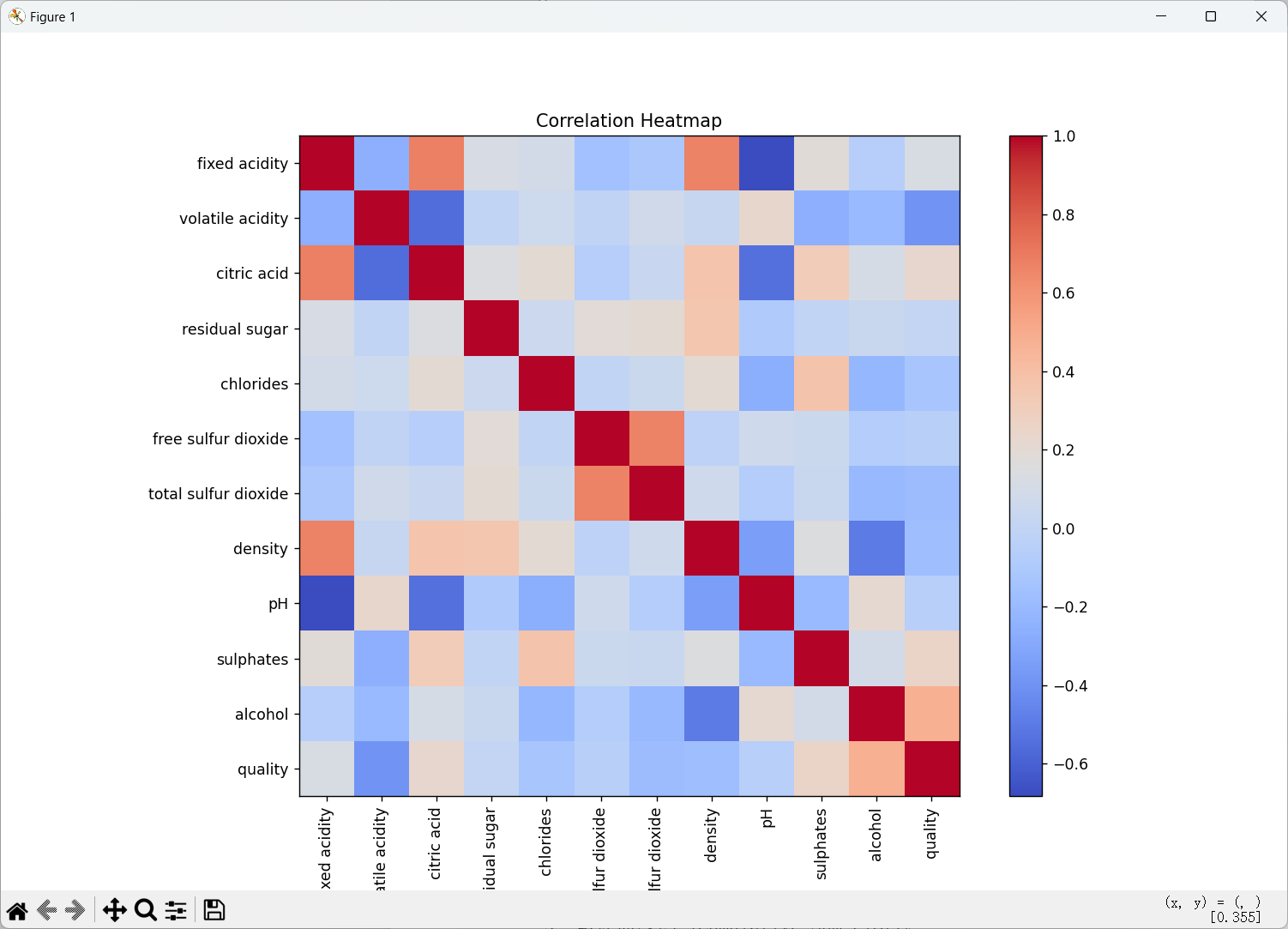
plt.show()











1. 箱线图 (Boxplot)

大多数字段的数据分布较为集中，但存在一些异常值。

residual sugar 和 total sulfur dioxide 字段的异常值较多，说明这些字段的数据分布较为分散。

quality 字段的分布较为均匀，没有明显的异常值。

2. 直方图 (Histogram)

大多数字段的数据分布呈现正态分布或接近正态分布。

residual sugar 和 total sulfur dioxide 字段的数据分布偏向右侧，说明大部分数据集中在较低的数值范围内，但存在一些较大的数值。

quality 字段的数据分布较为均匀，说明不同质量评分的红葡萄酒数量相对均衡。

3. 累计曲线图 (Cumulative Sum Plot)

大多数字段的累积和呈现线性增长，说明数据分布较为均匀。

residual sugar 和 total sulfur dioxide 字段的累积和增长较快，说明这些字段的数据分布较为分散，存在较大的数值。

4. 相关性热力图 (Correlation Heatmap)

fixed acidity 和 citric acid 之间存在较强的正相关性，说明固定酸度和柠檬酸含量之间存在一定的关系。

free sulfur dioxide 和 total sulfur dioxide 之间存在较强的正相关性，说明游离二氧化硫和总二氧化硫含量之间存在一定的关系。

density 和 residual sugar 之间存在较强的正相关性，说明密度和残糖含量之间存在一定的关系。

quality 与其他字段的相关性较弱，说明红葡萄酒的质量评分与其他化学成分之间的关系不明显。

5. 对角线矩阵的热力图 (Pairplot)

大多数字段之间的散点图呈现随机分布，说明字段之间的关系较弱。

fixed acidity 和 citric acid 之间的散点图呈现正相关，进一步验证了它们之间的关系。

free sulfur dioxide 和 total sulfur dioxide 之间的散点图呈现正相关，进一步验证了它们之间的关系。

1. 心得体会与建议

通过这次实验，我学会了如何使用Python进行数据的初步处理和可视化分析。通过箱线图、直方图、累计曲线图和热力图等多种可视化手段，我能够更直观地理解数据的分布和特征。这些图表帮助我发现了数据中的异常值和字段之间的相关性，为后续的深入分析和建模提供了重要的参考