计算机与信息工程学院实验报告

••••••••••••••••••••••••••••••••• 密 ••••••••••••••••••••••••••••••••• 封 ••••••••••••••••••••••••••••••••• 线 •••••••••••••••••••••••••••••••••

姓名：杨佳森\_学号：\_2112080106专业：数据科学与大数据技术\_年级：\_2021

课程： 数据分析和可视化 主讲教师：\_周黎鸣 辅导教师：\_无\_

实验时间： 2023 年\_6\_月7\_日\_下\_午10时至\_12\_时，实验地点

计算机大楼315\_\_\_\_\_

实验题目： 以“弘扬中国文化，坚定文化自信”为主题

实验目的： 根据所学知识，爬取和[我国文化](https://baike.baidu.com/item/%E4%B8%AD%E5%8D%8E%E6%96%87%E5%8C%96/2089243?fr=aladdin)相关的数据。利用数据预处理知识，分析数据利用可视化知识，展示数据，不少于5种类型的可视化展示。

实验环境（硬件和软件） jupyter

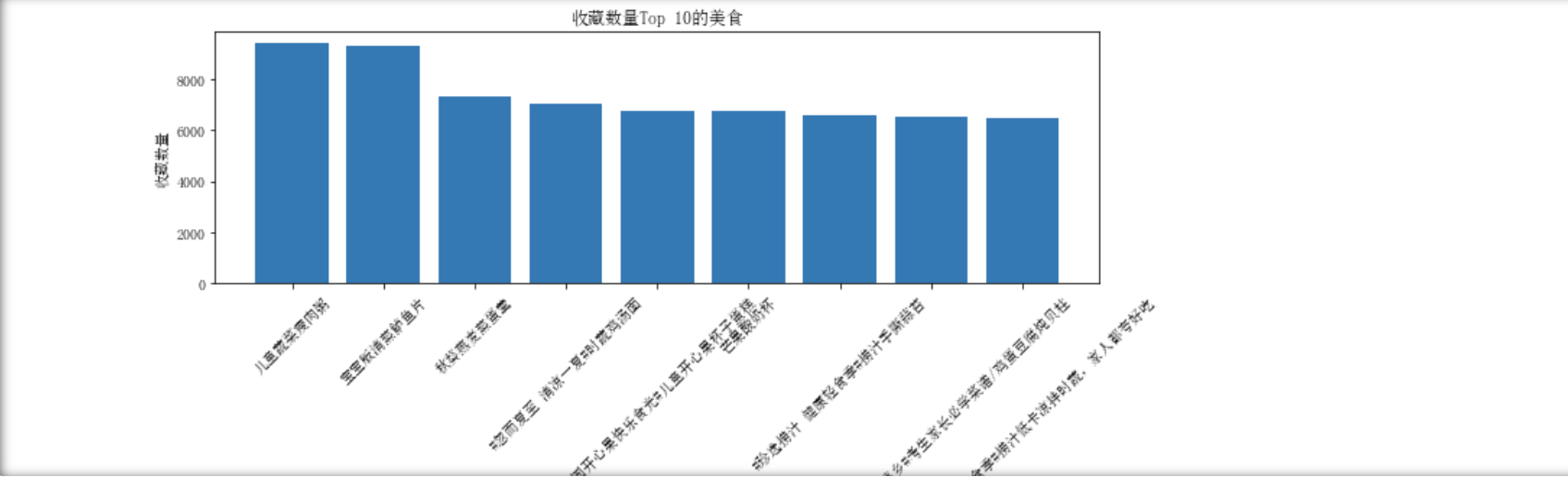
实验内容：

1. import numpy as np
2. import csv
3. import requests
4. from bs4 import BeautifulSoup
5. import pandas as pd
6. import matplotlib.pyplot as plt
7. from matplotlib.fontmanager import FontProperties
9. # 设置中文字符的字体
10. font = FontProperties(fname='C:/Windows/Fonts/simsun.ttc', size=12)
12. # 在绘制图形前设置字体
13. plt.rcParams['font.family'] = font.getname()

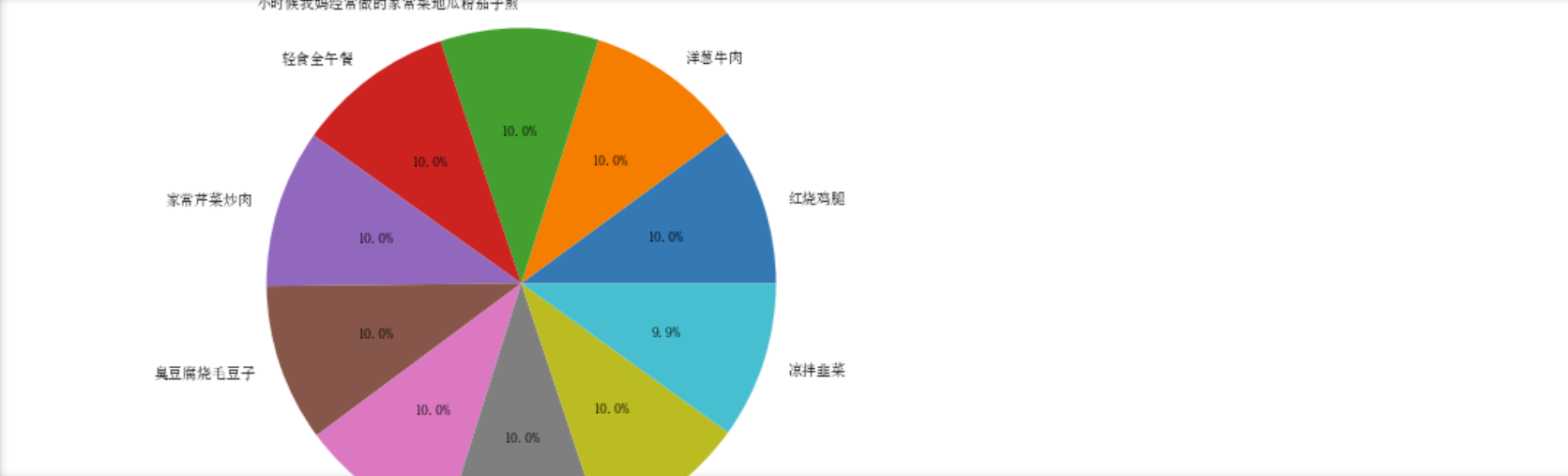
16. # 保存数据到CSV文件，采用utf-8编码，注意打开方式，否则会导致乱码
17. with open('recipes.csv', 'w', newline='', encoding='utf-8') as csvfile:
18. writer = csv.writer(csvfile)
19. writer.writerow(['标题', '作者', '浏览', '收藏'])
21. for i in range(63):
22. # 发起HTTP请求获取网页内容
23. url = 'https://www.douguo.com/jingxuan/{}'.format(i24)
24. headers={'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/114.0.0.0 Safari/537.36'}
25. response = requests.get(url, headers=headers)
26. soup = BeautifulSoup(response.text, 'html.parser')
28. # 提取美食相关信息
29. recipes = soup.findall('div', class='relative')
31. for recipe in recipes[:-1]:
32. title = recipe.find('a', class='cookname').text.strip()
33. author = recipe.find('a', class='author').text.strip()
34. view = recipe.find('span', class='view').text.strip()
35. collect = recipe.find('span', class='collect').text.strip()
36. writer.writerow([title, author, view, collect])
37. # 读取CSV文件为Dataframe，并更正数据类型
38. df = pd.readcsv('recipes.csv')
39. df['收藏'] = pd.tonumeric(df['收藏'], errors='coerce')
40. # 绘制条形图
41. top10collect = df.nlargest(10, '收藏')
42. plt.figure(figsize=(10, 6))
43. plt.bar(top10collect['标题'], top10collect['收藏'])
44. plt.xlabel('美食标题')
45. plt.ylabel('收藏数量')
46. plt.title('收藏数量Top 10的美食')
47. plt.xticks(rotation=45)
48. plt.tightlayout()
49. plt.show()
50. # 绘制饼图
51. df['浏览'] = df['浏览'].str.rstrip('万').astype(float) 10000
52. top10view = df.nlargest(10, '浏览')
53. plt.figure(figsize=(8, 8))
54. plt.pie(top10view['浏览'], labels=top10view['标题'], autopct='%1.1f%%')
55. plt.title('浏览量Top 10的美食')
56. plt.show()
57. # 绘制折线图
58. top100collect = df.nlargest(100, '收藏')
59. plt.figure(figsize=(12, 6))
60. plt.plot(top100collect.index, top100collect['收藏'], marker='o')
61. plt.xlabel('美食排名')
62. plt.ylabel('收藏数量')
63. plt.title('收藏数量Top 100的美食')
64. plt.xticks(rotation=45)
65. plt.tightlayout()
66. plt.show()
67. # 绘制箱线图
68. plt.figure(figsize=(8, 6))
69. plt.boxplot(top100collect['收藏'])
70. plt.ylabel('收藏数量')
71. plt.title('收藏数量Top 100的美食的箱线图')
72. plt.show()
73. # 绘制散点图
74. plt.figure(figsize=(8, 6))
75. plt.scatter(top100collect['浏览'], top100collect['收藏'])
76. plt.xlabel('浏览量')
77. plt.ylabel('收藏量')
78. plt.title('收藏量Top 100的美食中浏览量与收藏量的关系')
79. plt.show()

实验数据记录：

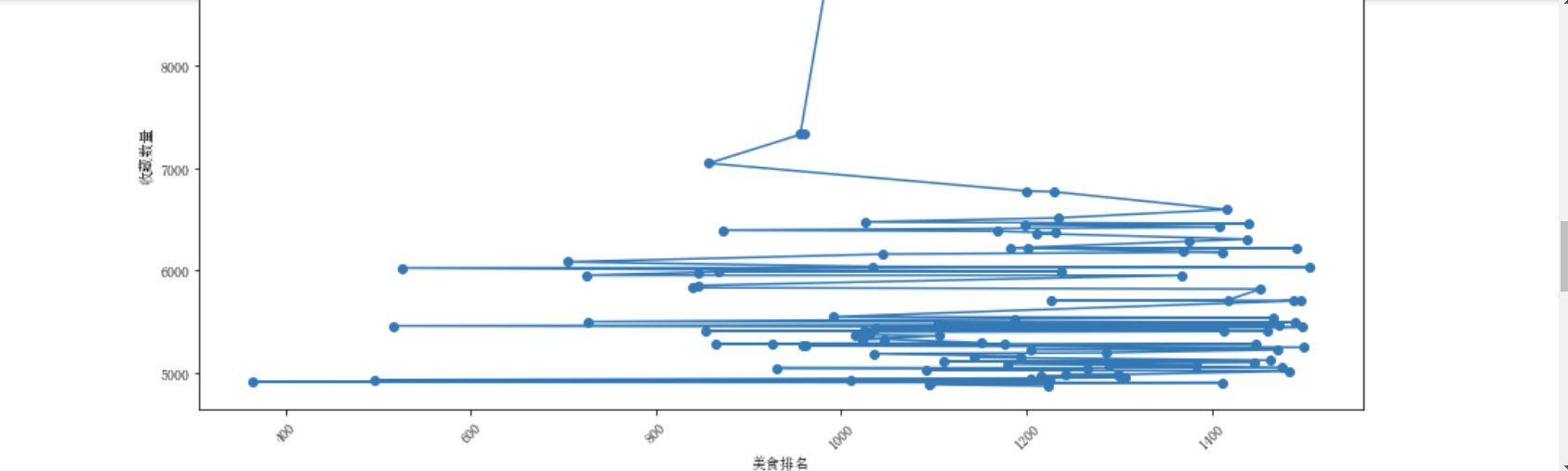
1、条形图（统计收藏数量Top 10的美食）



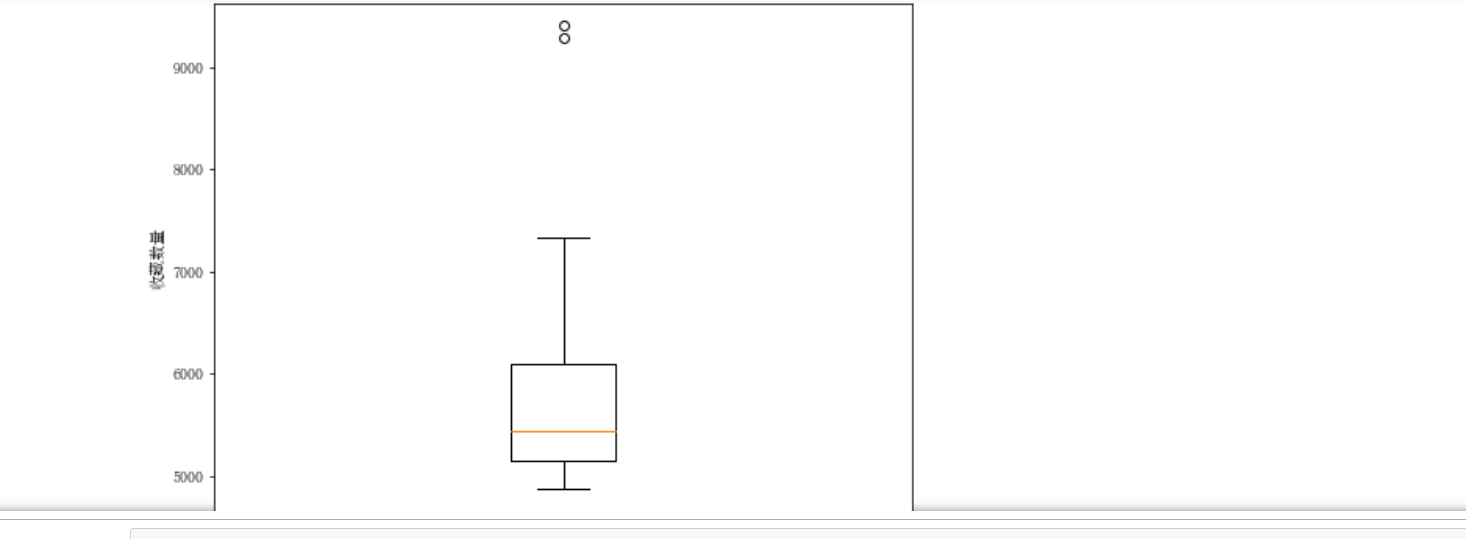
2、饼图（浏览量Top 10的美食）



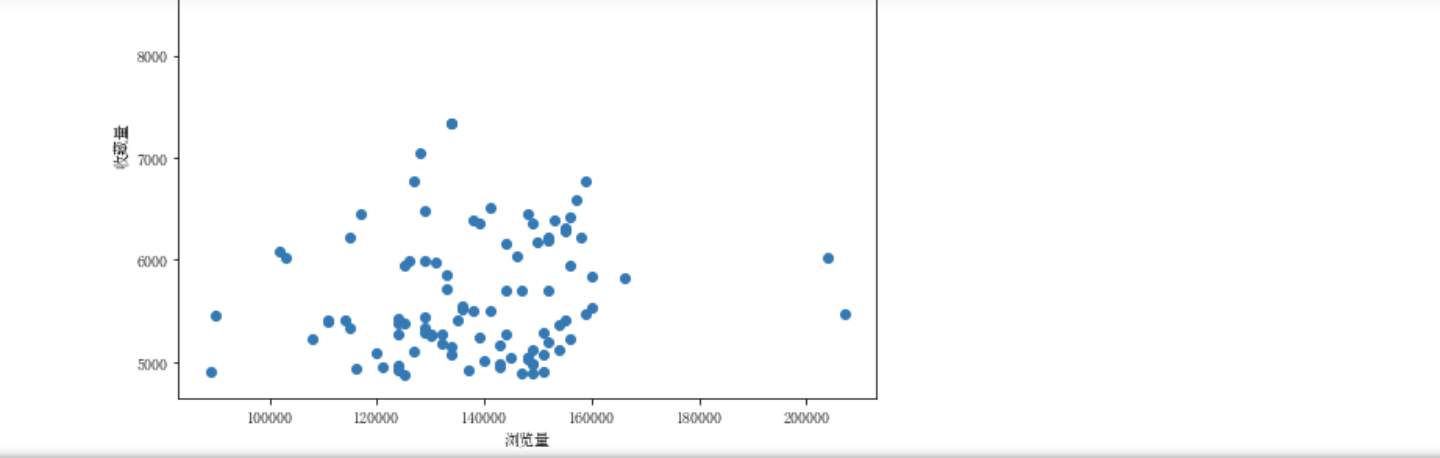
3、折线图（收藏数量Top 100的美食）



4、箱线图（收藏数量Top 100的美食的箱线图）



5、散点图（收藏量Top 100的美食中浏览量与收藏量的关系）



问题讨论：

豆果网站的菜谱菜单更新频率快，每次爬到的数据不会一样。

导出的csv文件用的是utf-8编码格式，用excel带开会乱码，注意打开方式。