

中

国

地

质

大

学

本科生课程报告

课程题目：

\*\*\*\*\*\*\*\*\*\*\*

课 程 名 称

Python 语言课程设计 A

刘远兴

教 师 姓 名

学 生 姓 名

\*\*\*

学 生 学 号

\*\*\*

日

期

2 0 \*\* 年 1 2 月 3 0 日

**目录**

[一、 题目要求 3](#_Toc27)

[1 、数据说明 4](#_Toc14174)

[2 、任务要求 4](#_Toc29215)

[二、需求分析 4](#_Toc2419)

[三、概要设计 4](#_Toc28964)

[四、详细设计 5](#_Toc25529)

[4.1主函数模块详细设计 5](#_Toc14063)

[4.2功能选择模块详细设计 5](#_Toc1184)

[4.3数据预处理和预览模块详细设计 7](#_Toc11524)

[4.4数据筛选及导出模块详细设计 8](#_Toc2338)

[4.5数据折线图示模块详细设计 8](#_Toc6099)

[4.6数据排序及导出模块详细设计 9](#_Toc22988)

[4.7数据统计及可视化模块详细设计 10](#_Toc1070)

[五、 上机操作 11](#_Toc17620)

[六、结果分析 14](#_Toc20850)

[七、致谢](#_Toc22040) [15](#_Toc22040)

八、参考文献................................................................15

## **题目要求**

1.1 数据说明

2006\_2010年美国五个著名城市纽约（New York）、华盛顿（Washington）、休斯顿（Houston）、洛杉矶(Los Angeles)、费城（Philadelphia）的空气中一氧化碳（NO）污染情况数据。

【字段说明】

序号（ID）：记录序号。

州代码（State Code）：由美国环保局分配给每个州的代码

县代码（County Code）：由美国环保署分配的特定州的代码

地点编号（Site Num）：由美国环保局分配的特定县的地点编号

地址（Address）：监测站点的地址

状态（State）：监测点的状态

县（Country）：县监测站点

城市（City）：监测点的城市

日期本地（Date Local）：监视日期

CO 污染物有 5 个专栏：

CO 单位（CO Unites）：测量 CO 的单位

CO 平均值（CO Mean）：给定日内 CO 浓度的算术平均值

CO 第一最大值（CO lst Max Value）：给定日期的 CO 浓度的最大值

CO 第一最大值时间：（CO 1st Max Hour）：给定日期的 CO 浓度的最大值所处的时间。

CO AQI（CO2 AQI）：一天内 CO 计算的空气质量指数

1.2 任务要求

【任务】

1、用pandas库读取“5. pollution\_us\_5city\_2006\_2010\_CO.csv”文件，查看前五行、后两行，并将缺失值全部丢弃处理。

2、选择City==“New York”、Date Local、CO Mean、CO 1st Max Hour四列导出为文本文件“pollution\_us\_NewYork\_2006\_2010\_COMean.txt”，要求数据之间用空格分隔，每行末尾包含换行符。

3、读取文本文件“pollution\_us\_ NewYork \_2006\_2010\_COMean.txt”，选择CO 1st Max Hour=“20”的所有行，以Date Local为横轴，以CO Mean为纵轴，画折线图。包括图例、图标题，x轴刻度以年显示，y轴显示刻度值，曲线颜色分别为红色。

4、选择City==“New York”、Date Local、CO AQI三列，按照列CO AQI 降序排序，并将排序后结果导出为Excel文件“pollution\_us\_NewYork\_2006\_2010\_COAQI.xlsx”。

5、读取Excel文件“pollution\_us\_NewYork\_2006\_2010\_COAQI.xlsx”，利用category = [0, 5, 10, 15，20，25，30]和labels = ['Good', 'Moderate', 'SubUnhealthy', 'Unhealthy', 'VeryUnhealthy', 'Hazardous']将CO AQI进行离散化，并根据离散化结果画出饼状图，保存为“CO\_AQI\_pie.png”，要求分辨率不低于300dpi。

【要求】

1 、根据以上数据处理任务，设计并编程实现“数据分析与可视化系统”， 要求

① 各个任务选择用菜单实现 (菜单可用字符串输出模拟，或者 Tkinter 形式 实现)。

② 各个任务名称自己定义，须由独立的函数实现，且每个任务执行成功与 否须给出必要的文字提示。

③ 数据输入和结果输出的文件名须由人工输入，且输出结果都要以文件形 式保存。

④ 为保持程序的健壮性，各个任务执行过程中需要进行必要的判断(如文 件是否存在、输入是否合法等)、程序异常控制等。

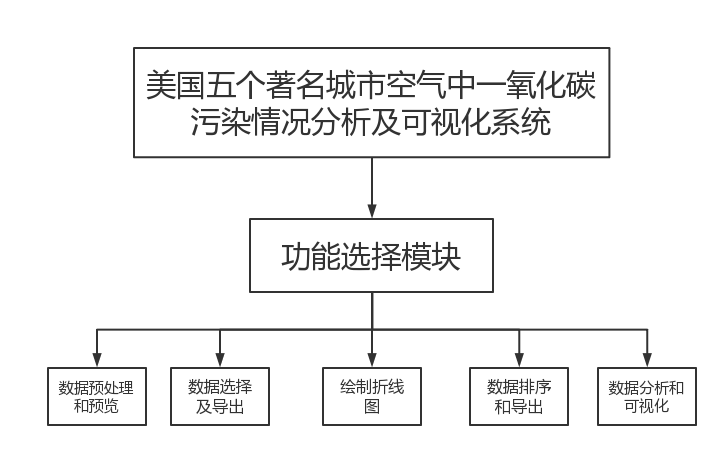
2 、根据以上统计结果，书写不少于 300 字的结果分析。

## 二、需求分析

根据题目要求，任务主要涉及数据读取、数据预处理、数据筛选、数据转存、数据统计、数据可视化、数据导出等常规的数据分析操作步骤，可以调用Pandas的文件读写、数据分析等功能模块实现；各任务要求用函数形式实现，需要设计各函数之间用参数传递实现各操作步骤之间的松耦合，进行模块化程序设计；各操作步骤需要用菜单实现功能选择，并提供必要的输入输出等人机交互操作；此外，在程序中应提供必要的异常控制代码，保证程序的健壮性。

三、概要设计

根据需求分析，可以将该系统设计为数据预处理及导出、数据选择及导出、绘制折线图、数据排序及导出、数据统计与可视化展示5大功能模块，以及功能选择主菜单辅助模块，如图1所示。



## 四、详细设计

### 4.1主函数模块详细设计

【分析】

在python函数式编程中，主函数一般比较简洁，只提供函数调用。在本例中，主函数仅包含任务调用函数。

【关键代码 】

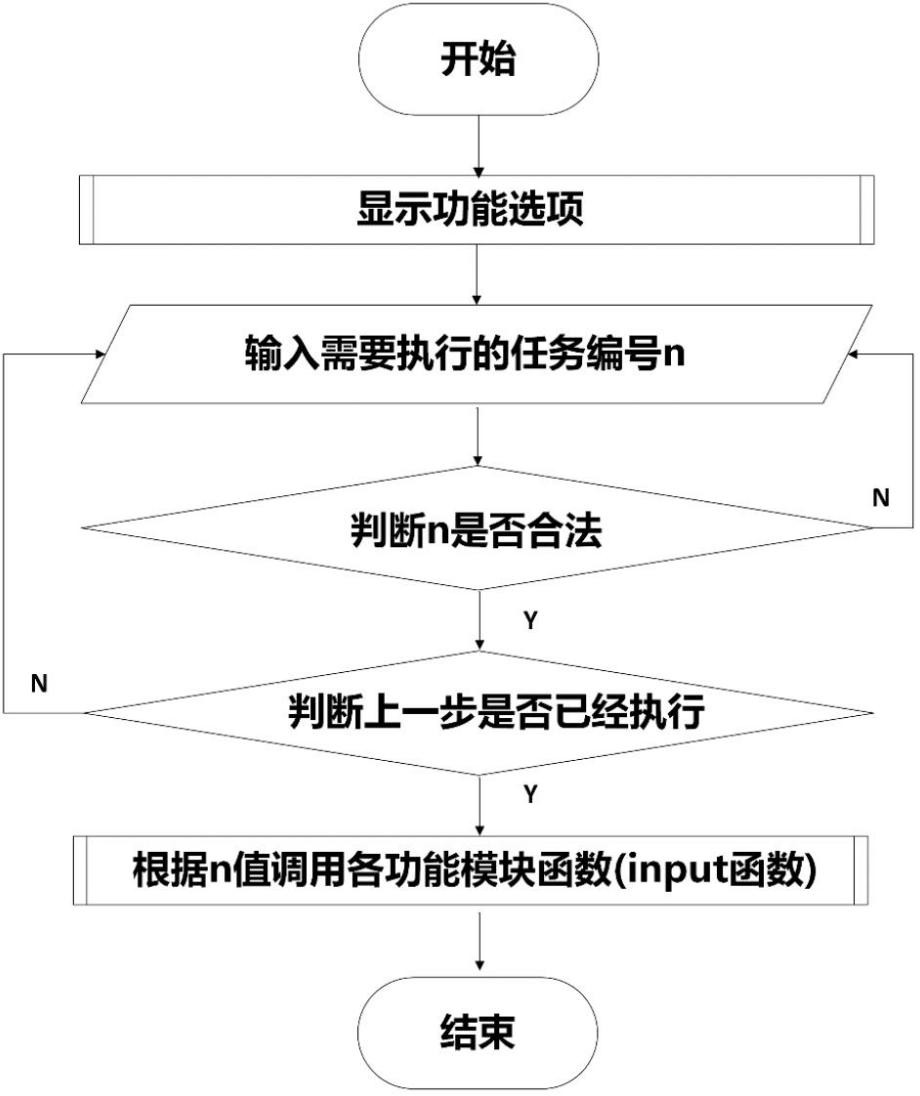
1. **if** \_\_name\_\_ == '\_\_main\_\_':
2. task()

### 4.2功能选择模块详细设计

【分析】

功能选择模块设计应比较简洁，只提供输入、功能处理、输出部分的函数调用，其中，各功能模块采用菜单时选择。另外，由于该实例中各任务间存在先后顺序关系，后续任务的输入可能来源于前序任务的输出，在各步骤执行前还需要判断需要的数据源是否已经准备好，即前序任务是否已经执行完毕。

具体流程如图2所示，图中N表示NO,Y表示YES。



【关键代码】

1. **def** menu():
2. **print**("【任务选择】\n"
3. '一一美国五个著名城市空气中一氧化碳污染情况分析及可视化系统－－－－－－一一\n'
4. ' | 0、退出。                         \n'
5. ' |1、数据预处理及预览。                        \n'
6. ' | 2、数据选择及导出。                        \n'
7. ' | 3、数据折线图。                        \n'
8. ' |4、数据排序及导出。                        \n'
9. ' |5、数据分析和可视化                   \n'
10. '+-----------------------------+')

13. **def** task():
14. **while** True:
15. menu()
16. num = input("请输入任务选项")
17. **if** num == '1':
18. dataPreprocessing()
19. **elif** num == '2':
20. **if** os.path.exists('5.pollution\_us\_5city\_2006\_2010\_CO.csv'):
21. dataSelection()
22. **else**:
23. **print**("未能执行当前选项，请执行先前选项")
24. **elif** num =='3':
25. **if** os.path.exists('pollution\_us\_NewYork\_2006\_2010\_COMean.txt'):
26. dataplot()
27. **else**:
28. **print**("未能执行当前选项，请执行先前选项")
29. **elif** num =='4':
30. **if** os.path.exists('pollution\_us\_NewYork\_2006\_2010\_COMean.txt'):
31. datarank()
32. **else**:
33. **print**("未能执行当前选项，请执行先前选项")
34. **elif** num == '5':
35. **if** os.path.exists('pollution\_us\_NewYork\_2006\_2010\_COAQI.xlsx'):
36. dataDescribeVisualization()
37. **else**:
38. **print**('未能执行当前选项，请执行之前的选项')
39. **elif** num == '0':
40. **print**('程序结束')
41. **break**
42. **else**:
43. **print**('输入程序有误')
44. input('回车显示菜单')

4.3数据预处理和预览模块详细设计

【分析】

数据预处理和预览模块是数据分析的重要组成部分，它主要包括数据清洗、数据转换、数据集成、数据规约和数据可视化等过程。以下是该模块的详细设计：数据清洗：该过程用于检测和纠正数据中的错误、缺失、重复和异常值。本示例要求将缺失值丢弃处理，因此本示例选择read\_csv来进行数据读取并且用对DataFrame对象的head（）和tail（）来查看前五行、后两行。并对数据预处理，丢弃废弃值

【关键代码】

1. **def** dataPreprocessing():
2. **while** True:
3. file\_name = input('请输入文件名5.pollution\_us\_5city\_2006\_2010\_CO.csv:')
4. **try**:
5. df = pd.read\_csv(file\_name,encoding='utf-8')
6. # 丢弃废弃值
7. df.dropna()
8. **print**("前五行")
9. # 前五行
10. **print**(df.head(5))
11. # 末两行
12. **print**('末两行')
13. **print**(df.tail(2))
14. **break**
15. **except**:
16. **print**("程序执行失败")

4.4数据筛选及导出模块详细设计

【分析】

数据筛选及导出模块是数据分析的重要组成部分，它主要包括数据筛选和数据导出两个过程。使用pandas库中的read\_csv函数读取第一个文件，并将其中'City'为'New York'的行的'Date Local'、'CO Mean'、'CO 1st Max Hour'三列提取出来，保存为一个新的dataframe对象df\_new。最后，将df\_new保存为一个以输入的第二个文件名命名的csv文件，

【关键代码】

1. **def** dataSelection():
2. **while** True:
3. **try**:
4. file\_name = input('请输入文件名5.pollution\_us\_5city\_2006\_2010\_CO.csv：')
5. file\_name0 = input('请输入要保存的文件名pollution\_us\_NewYork\_2006\_2010\_COMean.txt:')
6. df = pd.read\_csv(file\_name,encoding='utf-8')
7. # 选择‘City’=20的列以及其他列
8. df\_new = df.loc[df['City']=='New York',['Date Local','CO Mean','CO 1st Max Hour']]
9. df\_new.to\_csv(file\_name0,encoding='utf-8',index=None)
10. **print**('任务执行成功！')
11. **break**
12. **except**:
13. **print**('任务执行失败')

### 4.5数据折线图示模块详细设计

【分析】

该模块将部分数据用折线图呈现出来。使用pandas库中的read\_csv函数读取这个文件。接着，从读取的数据中选择CO 1st Max Hour=20的行，并将'Date Local'列的数据类型转换为日期类型。然后，将数据按照年份进行分组，计算每个年份的CO Mean的平均值，并将结果保存在一个新的dataframe对象df\_grouped中。接下来，使用matplotlib库绘制了一个折线图，横轴为年份，纵轴为CO Mean的平均值。

【关键代码】

1. **def** dataplot():
2. **while** True:
3. **try**:
4. file\_name = input('请输入文件名pollution\_us\_NewYork\_2006\_2010\_COMean.txt:')
5. df = pd.read\_csv(file\_name,encoding='utf-8')
6. # 行选择 CO 1st Max Hour=20 的行。
7. df\_new = df[df['CO 1st Max Hour'] == 20]
8. df\_new['Date Local'] = pd.to\_datetime(df\_new['Date Local'])
9. # 将 Date Local 列的数据类型转换为日期类型，以便后续按年份进行分组
10. df\_grouped = df\_new.groupby(df\_new['Date Local'].dt.year)['CO Mean'].mean()
11. # plt.figure(figsize=('10x10'))
12. #绘制折线图
13. plt.title('CO 1st Max Hour',fontsize = 16)
14. plt.xlabel('Date Local')
15. plt.ylabel('CO Mean')
16. plt.plot(df\_grouped.index, df\_grouped.values, color='red', label='CO Mean')
17. # x轴用年份表示
18. plt.xticks(df\_grouped.index, rotation=45)
19. # 设置图例
20. plt.legend()
21. plt.show()
22. **print**('任务执行成功！')
23. **break**
24. **except**:
25. **print**('任务执行失败')

### 4.6数据排序及导出模块详细设计

【分析】

数据排序及导出模块主要用于将数据按照指定的列进行排序，并将排序后的数据导出到指定的文件中。该模块可以帮助用户快速地对数据进行排序和导出，方便用户进行数据分析和决策。用pandas库中的read\_csv函数读取第一个文件，并将其中'City'为'New York'的行的'Date Local'、'CO AQI'两列提取出来，保存为一个新的dataframe对象df\_new。然后，按照'CO AQI'列的降序对df\_new进行排序。最后，将排序后的数据保存为一个以输入的第二个文件名命名的Excel文件，

【关键代码】

1. **def** datarank():
2. **while** True:
3. **try**:
4. file\_name = input('请输入文件名5.pollution\_us\_5city\_2006\_2010\_CO.csv：')
5. file\_name0 = input('请输入要保存的文件名pollution\_us\_NewYork\_2006\_2010\_COAQI.xlsx：')
6. df = pd.read\_csv(file\_name,encoding='utf-8')
7. # 筛选 City == 'New York'，并选择 Date Local 和 CO AQI 两列。
8. df\_new = df.loc[df['City']=='New York',['Date Local','CO AQI']]
9. # 按照 CO AQI 列的降序进行排序。
10. df\_new = df\_new.sort\_values(by='CO AQI',ascending= False)
11. df\_new.to\_excel(file\_name0,index=False)
12. **print**('任务执行成功')
13. **break**
14. **except**:
15. **print**('任务执行失败')

### 4.7数据统计及可视化模块详细设计

【分析】

数据统计，是数据分析的重要组成部分。

数据可视化，也是数据分析的重要组成部分。在 Python生态系统中,包括诸如matplotlib、Scaborm、HoloViews. Altair、PyQtGraph、ggplot、Bokch、pygal、VisPy、NetworkX、Plotly、geoplotlib、folium、Gleam、vincent、mpld3、python-igraph、missingno、Mayavi2、Leather等等众多的扩展库实现数据可视化功能。不同的可视化库在不同方面具有不同的偏重，如 Seaborn、Altair更偏重于专业统计图表绘制，PyQ1Graph、VisPy、Mayavi2更适合于数学、工程等领域制图，NctworkX.python-igraph 更适合于网络研究和分析制图geoplotlib、folium适合于绘制地图，等等。然而，在诸多的第三方可视化库中，matplotlib是Python中最为著名的绘图系统，很多其他的绘图系统如 seaborn 也是由其封装而来。此外，Pandas也提供了自己的绘图接口,并结合matlab实现更为丰富的功能。

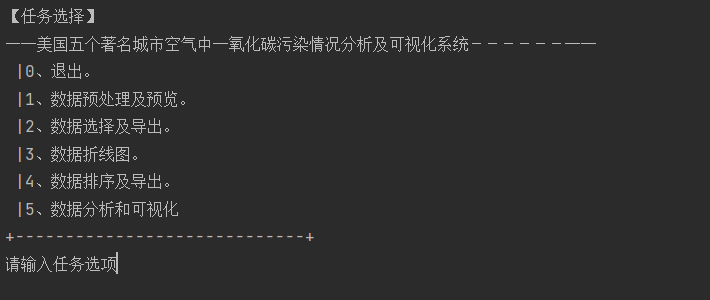
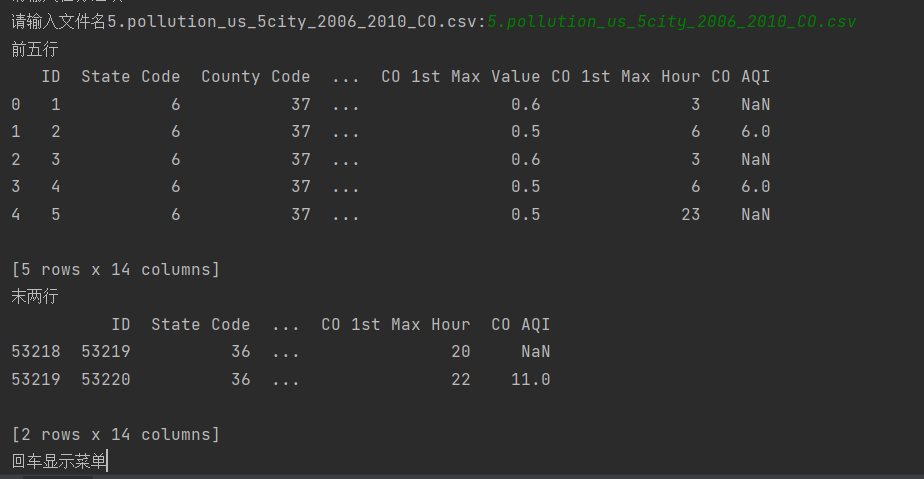
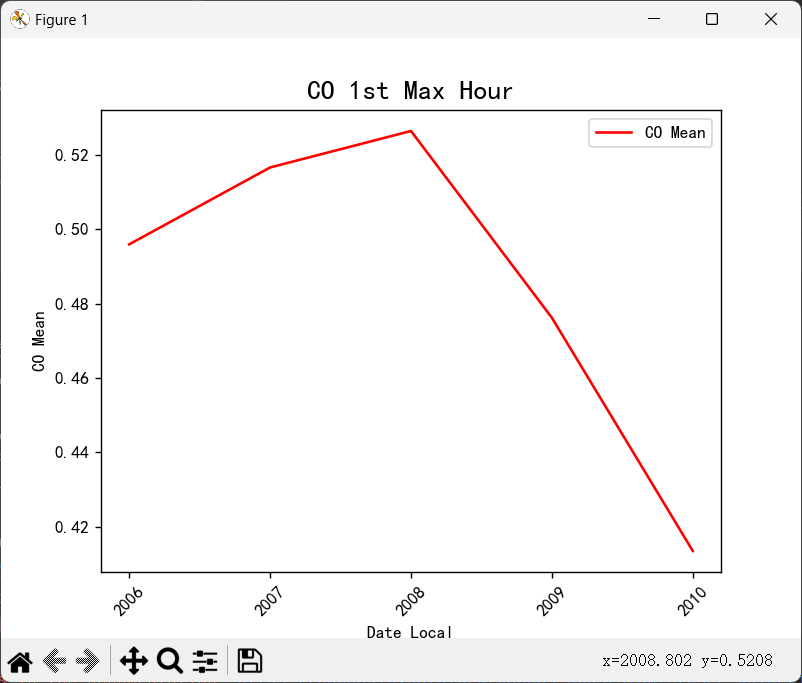
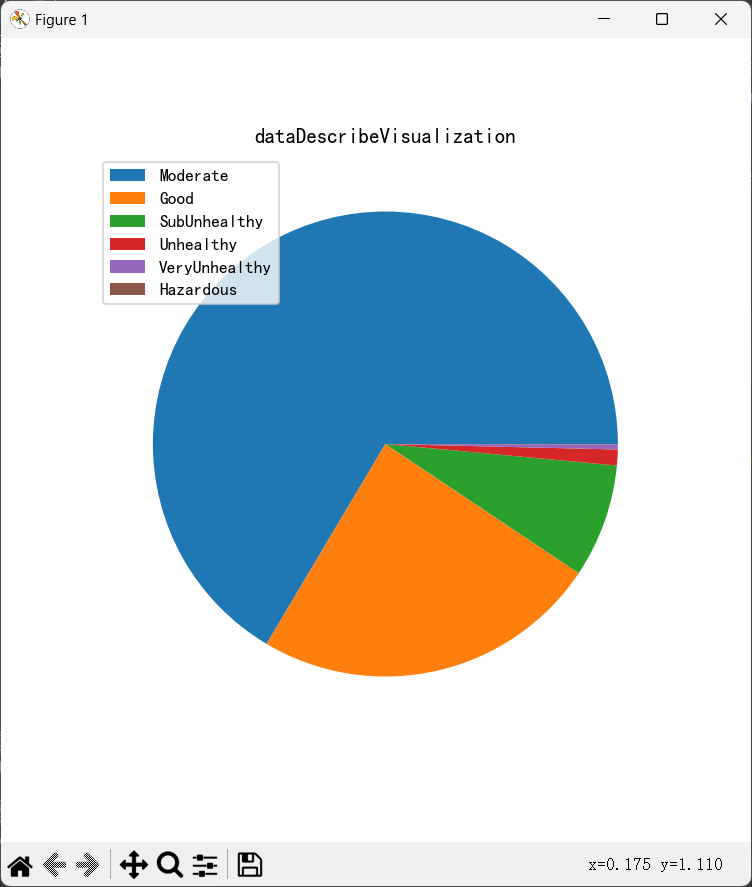
结合所学及题目所需，本题用pandas库中的read\_excel函数读取这个文件。接着，将'CO AQI'列的数据进行离散化处理，将其分为6个类别，并在dataframe对象df中添加一个新的列'CO AQI Category'来保存每个数据属于哪个类别。然后，统计每个类别的数量，并使用matplotlib库绘制了一个饼状图，用于展示每个离散化类别所占的比例。

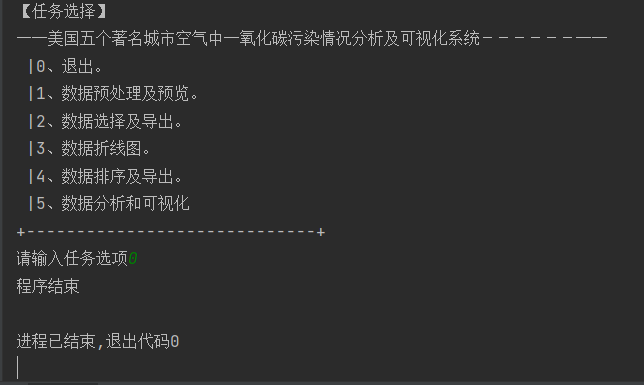
【关键代码】

1. **def** dataDescribeVisualization():
2. **while** True:
3. **try**:
4. file\_name = input('请输入文件名pollution\_us\_NewYork\_2006\_2010\_COAQI.xlsx：')
5. df = pd.read\_excel(file\_name)
6. # 将 CO AQI 列离散化
7. category = [0, 5, 10, 15, 20, 25, 30]
8. labels = ['Good', 'Moderate', 'SubUnhealthy', 'Unhealthy', 'VeryUnhealthy', 'Hazardous']
9. df['CO AQI Category'] = pd.cut(df['CO AQI'], bins=category, labels=labels)
10. # 统计每个离散化类别的数量
11. counts = df['CO AQI Category'].value\_counts()
12. # 绘制饼状图
13. plt.figure(figsize=(6,8))
14. plt.pie(counts.values, labels=None)
15. plt.title('dataDescribeVisualization')
16. # 添加图例，防止文字重叠
17. plt.legend(counts.index, loc=2)
18. plt.savefig('CO\_AQI\_pie.png',dpi = 300)
19. plt.show()
20. **print**('任务执行成功')
21. **break**
22. **except**:
23. **print**('任务执行失败')

## 上机操作

使用pycharm运行代码，实例如下：

1. 运行代码，进入主菜单页面
2. 输入‘1’，回车进入数据预处理及预览模块，输入文件名后执行
3. 输入‘2’，回车进入数据选择及导出模块，输入文件名后执行
4. 输入‘3’，回车进入数据折线图模块，输入文件名后执行，成功绘制折线图
5. 输入‘4’，回车进入数据排序及导出模块，输入文件名后执行，成功导出文件
6. 输入‘5’，回车进入数据分析和可视化模块，输入文件名后执行，成功绘制饼图。
7. 输入‘0’，回车退出程序。



## 六**、结果分析**

（1）污染水平：根据数据分析，五个著名城市在这段时间内的CO污染水平普遍较低。根据CO AQI（Air Quality Index）离散化结果的饼图显示，"Good"（优良）和"Moderate"（中等）的CO AQI分类占据了绝大部分比例。这表明这些城市的大气环境相对较好，CO污染水平处于可接受范围内。

（2）污染程度变化：在这五个城市中，不同年份CO污染程度可能存在变化。通过对数据进行更详细的分析，可以进一步了解每个城市的具体情况。例如，可以比较不同城市之间的CO AQI分类分布情况，以及每个城市随时间的CO AQI趋势。这样的分析将提供更多关于污染水平和可能的影响因素的洞察。

（3）空气质量改善：由于数据跨越了五年的时间范围，可以观察到空气质量改善的趋势。如果CO污染水平持续保持在良好和中等水平，并且没有明显的增加趋势，这表明城市在空气质量管理和监测方面采取了积极的措施。这种改善有可能归因于环保政策的实施、车辆排放标准的改善以及公众对环境问题的意识提高等因素。

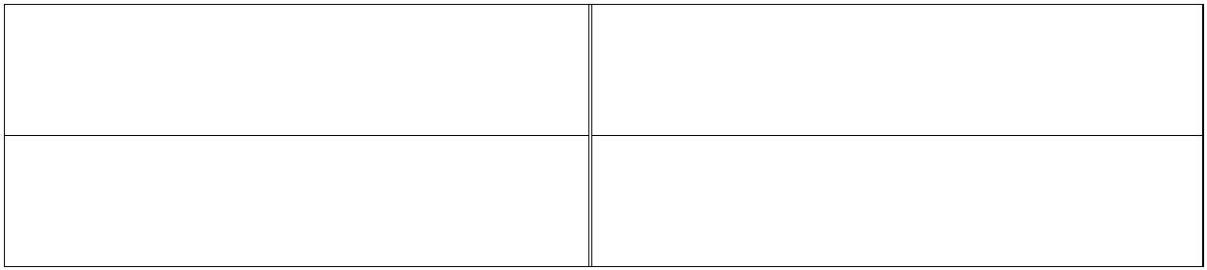
## 七、致谢

在此，我们要向周围的老师及同学致谢，他们对我们的帮助及引导。经过刘先生这一学期的指导，不仅使我对 Python的基本理论有了更深的认识，而且使我的程序设计技能得到了很大的提高，同时也使我对一些具体的问题有了更深刻的认识。另外，我在数据的分析和程序的功能的实现上还有许多的缺陷，我希望在未来的日子里，我可以对 Python的爬行技术进行更深的研究，逐渐地完善自己的知识体系，并积累更多的实践经验。

1. **参考文献**

[1]闫继宁，陈云亮，王媛妮，宋维静．Python 语言应用案例实践教程［Ｍ].科学

出版社，2021 年 5 月．



注：1、无评阅人签名成绩无效；

2、必须用钢笔或圆珠笔批阅，用铅笔阅卷无效；

3、如有平时成绩，必须在上面评分表中标出，并计算入总成绩。

平时成绩：

课程论文成绩：

总 成 绩：

评阅人签名：

课程设计评语

对课程论文的评语: