

山东建筑大学

Shandong Jianzhu University

数据可视化技术

实验指导书

(适用专业：计算机科学与技术)

山东建筑大学计算机学院

计算机科学与技术教研室

2021.4

目录

使用说明.....	3
实验一 数据导入与数据处理	4
实验二 类别型数据可视化	9
实验三 数据变量间关系可视化.....	10
实验四 数据分布特征可视化	12
实验五 时间序列可视化.....	14
实验六 局部整体型数据可视化.....	15
实验七 高维数据可视化基础	16
实验八 地理空间数据可视化基础	18

使用说明

1. 关于实验原理

实验原理已在上课时讲述，并给出了实例代码，故在本实验指导书中不再重复撰写原理部分，仅给出实验任务，每一部分的实现请参考授课代码。

每部分实验可以在授课代码基础上进行修改获得实验结果，如果给定了多种实现方式代码，至少要实现一种，多多益善。

学生可以自行用授课以外的其他库完成任务，实验结果评定等同对待，鼓励学生使用其他可视化库实现。

2. 关于实验报告

1) 实验报告撰写请采用教师给定的模板撰写。

2) 实验报告内容包括实验目的、实验环境、实验内容和实验总结四项，缺一项在实验报告评定时降低一档评定。实验目的和实验环境请按实验指导书撰写即可。实验内容部分，对每个小任务，给出实现代码，并将实验结果截图贴在下面。实验总结主要撰写在本次实验中遇到的问题，用什么方法解决的做个记录，同时对实践内容有什么感悟或领悟，也可以写在里面。

3) 实验报告只收电子稿，学生需要将其转换为 PDF 格式，通过作业系统上传，请勿提交其他格式，因格式提交不正确造成成绩评定异常的责任自负。

实验一 数据导入与数据处理

实验目的

1. 了解常用的数据存储格式
2. 掌握从数据文件导入数据的方法
3. 掌握从数据库导入数据的方法
4. 掌握Pandas的DataFrame和Series数据类型基本操作

实验环境

Windows10, Anaconda3-2020.11-Windows-x86_64, Jupyter Lab, MySQL

实验内容

1. 数据获取

实验前请下载数据集文件: orders.txt, orders.csv, orders.xls, orders.xlsx, 安装好MySQL数据库或其他数据, 若使用MySQL数据库, 请安装pymysql库, 若使用SQLServer, 请安装pymssql库, 其他数据请安装相应的python访问数据库的包。请将order.xls导入到数据库中。进行以下实验, 并显示每个结果的前5行数据。

- 1) 从txt格式文件读取数据
- 2) 从csv格式文件读取数据
- 3) 从xls、xlsx格式文件读取数据
- 4) 从JSON格式文件读取数据
- 5) 从关系型数据库获取数据

2. Series数据的基本操作

有个学生名字为张三, 学号20180015, 英语、程序设计、高数、物理成绩分别为80、90、95、88, 请采用上述数据进行下列实验, 操作完成后显示结果。

2.1 Series数据的创建

- 1) 通过列表创建Series, 并指定索引

先建立表述学者信息的列表, 接着使用 “姓名、学号、英语、程序设计、高数、物理” 作为索引把列表转换为Series。

- 2) 通过字典创建Series

先建立表述学生信息的字典，接着使用 “姓名、学号、英语、程序设计、高数、物理” 作为索引把列表转换为Series。

2.2 Series数据的索引

使用 “高数” 作为索引值，取出高数课程成绩并显示。

2.3 Series数据的修改

使用 “物理” 作为索引值，修改物理课程成绩为80。

2.4 Series数据的运算

在2.3创建的Series基础上，添加一个元素 “总分”，并将4门课程总分做该元素的值。

3. DataFrame的基本操作

某大型设备制造企业有 A、B、C3 种型号的大型设备，在 2010 年三种设备分布卖出了 1 台、3 台、4 台，在 2011 年分别卖出了 3 台、5 台、2 台。请根据下列要求进行操作。

3.1 数据框的创建

1) 由列表创建数据框

创建列表，用于表示每种设备每年销售情况，通过列表使用 DataFrame 函数建立数据框，制定列名为 ‘X’、‘year’、‘value’。

2) 使用嵌套字典创建数据框

创建嵌套字典，通过嵌套字典使用 DataFrame 函数建立同 1) 中描述相同的数据框。

3.2 数据框的选取

在第 3.1 部分基础上完成以下操作：

- 1) 选取 “value” 列
- 2) 选取 “year” 和 “value” 列两个列
- 3) 选取首行数据
- 4) 选取第 0 行和第 1 行
- 5) 选取第 3 行第 2 个元素
- 6) 选取销量大于等于 3 的数据

3.3 数据框的多重索引

在第 3.1 部分基础上完成以下操作：

1) 设置多重索引

将“X”和“year”两列设置为索引，建立两重索引，其中第 0 级索引为“X”，第 1 级索引为“year”。

2) 使用多重索引定位数据

使用双重索引定位“value”中首个元素。

3.4 数据框的表格变换

在第 3.1 部分基础上完成以下操作：

1) 将长数据转换为宽数据

用 X 列做索引，按 year 列转换为宽数据。

2) 将宽数据转换为长数据

将 1) 中宽数据转换为长数据。

3.5 数据框的排序

在第 3.1 部分基础上完成以下操作：

使用数据框的 `sort_value()` 方法，采用“value”做关键字，对数据框分别进行升序排序和降序排序。

3.6 数据框的拼接

x	y
a	1
b	2
c	3

z	g
B	2
D	5
H	3

x	y
g	2
d	5

图 1-1 数据框的示意图

创建如图 1-1 所示的三个数据框，进行以下实验：

1) 沿横轴方向拼接 df1 和 df2

2) 沿纵轴方向拼接 df1 和 df3

3) 删除数据框 df1 的 ‘y’ 列

4) 删除数据框 df1 的第 1 行

3.7 数据框的融合

df1

x	y
a	1
b	2
c	3

df2

x	z
a	2
b	5
d	3

df4

x	y	z
a	1	2
b	4	5
d	2	3

图 1-2 数据框的示意图

创建如图 1-2 所示的三个数据框，使用 merge() 函数进行以下实验，

- 1) 只保留左表的所有数据
共同列选'x'列，设置 left=df1, right=df2, 进行实验。
- 2) 只保留右表的所有数据
共同列选'x'列，设置 left=df1, right=df2, 进行实验。
- 3) 保留两个表的公共部分信息
共同列选'x'列，设置 left=df1, right=df2, 进行实验。
- 4) 保留两个表的所有信息
共同列选'x'列，设置 left=df1, right=df2, 进行实验。
- 5) 多列匹配
共同列选'x'列和'y'列，设置 left=df1, right=df4, 进行实验。

3.8 数据框的分组操作

构建如图 1-3 所示数据框，采用该数据框进行如下实验：

	x	2010	2011
0	A	1	3
1	B	3	5
2	C	4	2
3	A	4	8
4	C	3	9

图 1-3 数据框结构

- 1) 按行求和

指定对“2010”和“2011”两列进行运算，求每一行的和。

2) 按列求和

指定对“2010”和“2011”两列进行运算，求每一行的和。

3) 单列运算

对数据框添加一列，名为‘2010_2’，数据为‘2010’列数据加 2。

4) 多列运算

对数据框添加一列，名为‘2010_2011’，数据为‘2010’列数据乘 2 加上‘2011 列的数据’。

5) 使用 groupby 函数进行分组操作

按‘year’分组求和。

6) 使用 groupby 函数进行分组操作

按‘year’分组求方差。

实验二 类别型数据可视化

实验目的

1. 了解类别型数据的形式；
2. 掌握柱状图、条形图、雷达图、玫瑰图的编程绘制方法
3. 熟悉类别型数据的可视化分析

实验环境

Windows10, Anaconda3-2020.11-Windows-x86_64, Jupyter Lab

实验内容

本次实验采用含有2000个记录的网购满意度情况调查数据，进行实验。实验用数据集的前5行数据如图2-1所示。本数据集有3列，均为类别型数据。

性别	网购次数	满意度
女	6次以上	满意
男	3-5次	不满意
女	3-5次	不满意
男	3-5次	中立

图2-1 实验数据集概览

1. 柱状图、条形图绘制

1.1 单数据系列柱形图

按照“性别”列，统计参与网购的男、女购物次数，绘制柱形图。

1.2 单数据系列条形图

按照“满意度”列，统计参与网购的满意情况次数，绘制条形图。

1.3 多数据系列条形图

按照“性别”和“网购次数”两个列，分别统计不同性别，不同网购次数的购物次数，以“网购次数”做类别横轴，绘制多数据系列条形图。

2. 绘制雷达图

按照“性别”和“网购次数”两个列，分别统计不同性别，不同网购次数的购物次数，按照男、女两个大类，绘制网购次数雷达图。

3. 绘制玫瑰图

按照“满意度”列，统计参与网购的满意情况次数，绘制玫瑰图。

实验三 数据变量间关系可视化

实验目的

1. 了解数据变量间关系可视化所需数据框的结构形式
2. 掌握散点图、气泡图、相关系数图的编程方法
3. 熟悉用于数据间关系的可视化图表的应用特点

实验环境

Windows10, Anaconda3-2020.11-Windows-x86_64, Jupyter Lab

实验内容

本实验采用上市公司的各项财务指标数据集，在创业板、中小板和主板中随机抽取地产类、医药类、科技类和食品类股票共200只，该数据集具有如图3-1所示的形式。

	股票类型	上市板块	总股本	每股收益	每股净资产	净资产收益率	每股资本公积金	每股现金流量
0	医药类	创业板	32062	1.82	11.21	8.11	2.98	3.42
1	地产类	中小板	53697	0.70	3.96	4.76	0.48	2.99
2	地产类	中小板	81757	0.94	7.85	6.32	2.19	3.05
3	医药类	创业板	27392	1.95	10.24	7.99	3.03	3.36
4	科技类	中小板	48395	0.90	11.52	7.59	4.65	3.56

图3-1 上市公司财务数据数据集概览

1. 散点图绘制

绘制每股收益与每股净资产之间的关系，绘制散点图，分析二者之间是正相关关系还是负相关关系，或不相关。

2. 绘制带趋势曲线的二维散点图

以总股本和净资产收益率两列数据的散点图，并采用LOESS方法绘制数据平滑曲线。

3. 绘制相关系数图

从数据集中筛选出创业板的股票数据，绘制总股本、每股收益、每股净资产、净资产收益率、每股资本公积金、每股现金流之间的相关系数图，以热力图和方块图进行可视化展示。

4. 绘制气泡图

从数据集中筛选出医药类股票数据，以总股本、每股收益绘制气泡图，气泡大小展示净资产收益率。

实验四 数据分布特征可视化

实验目的

1. 了解数据特征分布的基本概念
2. 掌握一维数据统计直方图和核密度图、抖动散点图、箱线图、小提琴图和二维统计直方图和核密度图的编程绘制方法

实验环境

Windows10, Anaconda3-2020.11-Windows-x86_64, Jupyter Lab

实验内容

本实验采用北京市2018年全年的空气质量数据集。数据集的形式如图4-1所示。试根据该数据集做如下分析：

	日期	AQI	质量等级	PM2.5	PM10	二氧化硫	一氧化碳	二氧化氮	臭氧浓度
0	2016-01-01	226	重度污染	176	199	33	3.4	106	12
1	2016-01-02	316	严重污染	266	299	35	4.6	124	16
2	2016-01-03	297	重度污染	247	296	20	3.5	83	22
3	2016-01-04	53	良	37	23	6	0.9	33	59
4	2016-01-05	63	良	34	76	11	0.9	38	50

图4-1北京市空气质量数据集概览

1. 直方图和核密度图的绘制

从数据集中提取空气质量指数（AQI）数据，绘制该指标的统计直方图和核密度图。

2. 绘制堆叠统计直方图

从数据集中提取空气质量指数（AQI）数据和质量等级数据，按照空气质量等级类别，绘制堆叠统计直方图。

3. 绘制抖动散点图

从数据集中提取臭氧浓度数据，绘制抖动散点图。

4. 绘制箱线图和小提琴图

从数据集中提取PM2.5浓度、PM10、二氧化硫、一氧化碳、二氧化氮和臭氧浓度数据，分别绘制箱线图和小提琴图。

5. 绘制二维统计直方图和核密度图

从数据集提取PM2.5和PM10浓度数据，绘制二维统计直方图和核密度图。

实验五 时间序列可视化

实验目的

1. 了解时间序列数据的意义
2. 掌握编程实现折线图、面积图、日历图的方法

实验环境

Windows10, Anaconda3-2020.11-Windows-x86_64, Jupyter Lab

实验内容

本实验采用北京市2018年全年的空气质量数据集。数据集的形式如图5-1所示。试根据该数据集做如下分析：

	日期	AQI	质量等级	PM2.5	PM10	二氧化硫	一氧化碳	二氧化氮	臭氧浓度
0	2016-01-01	226	重度污染	176	199	33	3.4	106	12
1	2016-01-02	316	严重污染	266	299	35	4.6	124	16
2	2016-01-03	297	重度污染	247	296	20	3.5	83	22
3	2016-01-04	53	良	37	23	6	0.9	33	59
4	2016-01-05	63	良	34	76	11	0.9	38	50

图4-1北京市空气质量数据集概览

1. 绘制折线图

从数据集中提取PM2.5浓度数据，绘制北京市2018年PM2.5浓度折线图。

2. 绘制面积图

从数据集中提取空气质量指数（AQI）8月份的数据，绘制面积图。

3. 绘制日历图

从数据集中提取一氧化碳浓度数据，按月份绘制一氧化碳浓度日历图。

实验六 局部整体型数据可视化

实验目的

1. 了解局部整体型数据可视化的数据特点
2. 掌握饼状图、圆环图、分裂饼状图、内嵌环形饼图的绘制编程方法

实验环境

Windows10, Anaconda3-2020.11-Windows-x86_64, Jupyter Lab

实验内容

本次实验采用含有2000个记录的网购满意度情况调查数据，进行实验。实验用数据集的前5行数据如图2-1所示。本数据集有3列，均为类别型数据。请依据该数据集做以下工作：

性别	网购次数	满意度
女	6次以上	满意
男	3-5次	不满意
女	3-5次	不满意
男	3-5次	中立

图6-1 实验数据集概览

1. 绘制饼状图

从数据集中提取满意度数据，统计网购持满意、不满意、中立态度的次数，绘制饼状图。

2. 绘制分裂饼图

在第一部分基础上绘制分裂饼图

3. 绘制圆环图

从数据集中提取网购次数数据，统计计算网购次数1-2次、3-5次和6次以上人数，依据人数绘制圆环图。

4. 绘制内嵌环形饼图

从数据集中提取性别和满意度数据，分别统计男人和女人对网购持满意、不满意、中立态度的次数，绘制内嵌环形饼图，对比男人和女人的满意度情况。

实验七 高维数据可视化基础

实验目的

1. 了解高维数据可视化所需要的数据处理方法
2. 掌握分面图、矩阵散点图、平行坐标图的编程实现方法

实验环境

Windows10, Anaconda3-2020.11-Windows-x86_64, Jupyter Lab

实验内容

本实验采用上市公司的各项财务指标数据集，在创业板、中小板和主板中随机抽取地产类、医药类、科技类和食品类股票共200只，该数据集具有如图7-1所示的形式。

	股票类型	上市板块	总股本	每股收益	每股净资产	净资产收益率	每股资本公积金	每股现金流量
0	医药类	创业板	32062	1.82	11.21	8.11	2.98	3.42
1	地产类	中小板	53697	0.70	3.96	4.76	0.48	2.99
2	地产类	中小板	81757	0.94	7.85	6.32	2.19	3.05
3	医药类	创业板	27392	1.95	10.24	7.99	3.03	3.36
4	科技类	中小板	48395	0.90	11.52	7.59	4.65	3.56

图7-1 上市公司财务数据集概览

1. 绘制分面图

1.1 绘制3个维度分面图

从数据集中选取上市板块、总股本和每股收益数据，分别绘制创业板、中小板和主板上市企业每股收益和总股本的散点图。

1.2 绘制4个维度分面图

从数据集中选取上市板块、总股本、每股收益和净资产收益率数据，分别绘制创业板、中小板和主板上市企业每股收益和总股本的散点图，点的按照每股净资产收益率数据确定。

1.3 绘制5个维度分面图

从数据集中选取上市公司的股票类型、上市板块、每股收益、总股本、净资产收益率等数据，绘制矩阵分面气泡图。

2. 绘制矩阵散点图

从数据集中选取上市公司的每股收益率、总股本、净资产收益率数据，绘制矩阵散点图，对角线上采用统计直方图。

3. 绘制平行坐标图

从数据集中提取除股票类型外的所有数据，按照上市板块分类，绘制平行坐标图。

实验八 地理空间数据可视化基础

实验目的

- 1.了解数据可视化所需地理信息数据存储形式；
- 2.掌握分级统计地图、点描法地图编程实现方法

实验环境

Windows10, Anaconda3-2020.11-Windows-x86_64, Jupyter Lab

实验内容

本实验使用的山东省地图数据以SHP文件形式给定；山东省17地市（莱芜在地图数据中依然是独立的）人口与GDP数据集如图8-1所示。依据给定的数据进行以下实验：

	NAME99	人口	总面积	GDP排行	2019年GDP	增速
0	滨州市	392.25	9453.00	13	2457.19	4.1
1	德州市	581.00	10356.00	9	3022.27	6.1
2	东营市	217.21	8243.00	11	2916.19	4.2
3	菏泽市	876.50	12238.62	8	3409.98	6.3
4	济南市	746.04	10244.00	2	9443.37	7.0

图8-1山东省17地市GDP数据集概览

1.绘制地图

读取给定的SHP格式的山东省地图数据，绘制山东省机场地图。

2. 绘制分级统计地图

根据山东省17地市数据，以人口数据做统计数据，在山东省基础地图的基础上绘制分级统计地图。

3. 点描法绘制带气泡的地图

根据山东省 17 地市数据，以人口数据做控制气泡大小，在山东省基础地图的基础上点描法绘制带气泡的地图。