



中国石油大学 (华东)
CHINA UNIVERSITY OF PETROLEUM

2021—2022 学年第 2 学期 《大学计算机》课程报告

选题名称		京东商城 ‘AJ1 休闲鞋’ 价格等数据处理分析		
小组成员	学号	姓名	任务分工	备注
	2001030118	王岩	数据爬取处理、整理编写报告	
	2001030104	郭文静	数据分析处理	
	2001030123	张成业	数据可视化处理	
评价指标				教师评分
1) 文档：结构完整，有条理；格式规范，排版好；语言通顺，错误少（20%） 2) 数据获取与清洗：过程清晰，方法得当，描述清楚准确（30%） 3) 数据处理、数据可视化：图表选择合理，方法恰当，描述清晰；界面美观、效果好；数据处理方法科学有效，描述完整、清晰（30%） 4) 数据分析：分析合理，逻辑性好；结论、观点有说服力（20%）				
教师评语				
教师签名： 2022 年 月 日				

说 明

1. 本课程要求学生把自己或小组实现的实验结果展示出来，重点描述对某个领域数据进行获取、分析、可视化的方法和过程，形成分析结果和结论。因此，课程报告主要包括任务要求、数据获取、分析、可视化展示等数据处理过程的内容。
2. 课程结束后，需要提交课程报告（Word 文档）、源程序和数据文件（WinRAR 压缩包）、演示文稿（PowerPoint 文档）和 5 分钟短视频（MP4 文件）。未提交相关资料者不能参加成绩评定。
3. 课程报告的段落结构和格式要求参考本文档（可作为模板），可以做适当的调整，文字内容需自己撰写。文档结构要完整、格式排版应美观、条理要清晰、论述需准确。
4. 抄袭、剽窃他人作品，记零分。

注意事项

- （1）报告需保留该模板的首页和本页内容，首页中的“成员”部分（包括任务分工）需要学生自己填写，并改为正常字体。评分、评语由教师填写。
- （2）撰写实验报告时，请删除该模板中用红色字体（正体或斜体）标注的文字，比如这一段。模板中的蓝色斜体字部分是需要自己编写的内容，包括图、表、公式、代码等。成文时，应该为标准格式，比如黑色、正体等等

京东商城 ‘AJ1 休闲鞋’ 价格等数据处理分析

一、任务描述

2021 年，中国网络零售额达 13.1 万亿元，创历史新高。以京东、天猫、苏宁易购为首的互联网企业是网购发展的核心力量。

如此庞大的经济产量来源于每个民众对网络购物的使用贡献。选购商品时，网站平台会把多件相似商品罗列到网页，并且直观地展示每一件商品的样式、价格、名称等信息以供消费者挑选（如图 1）。面对品目繁多的商品，大部分消费者都会‘看花眼’，这也无形之中为消费者增添了些许烦恼。如果能够把网站所提供商品的各类信息进行分类，消费者根据自己的需求及分类信息进行挑选便会事半功倍。

本课题旨在利用程序设计 python 语言爬取京东网站一特定商品（以 AJ1 休闲鞋为例（如图 2））价格等信息，并将信息存储至 excel 文件；对各类商品信息进行分析、提取与整合，最后绘制各类图表直观地展示多件商品的分类信息，为不同需求的消费者提供简便的选择渠道。

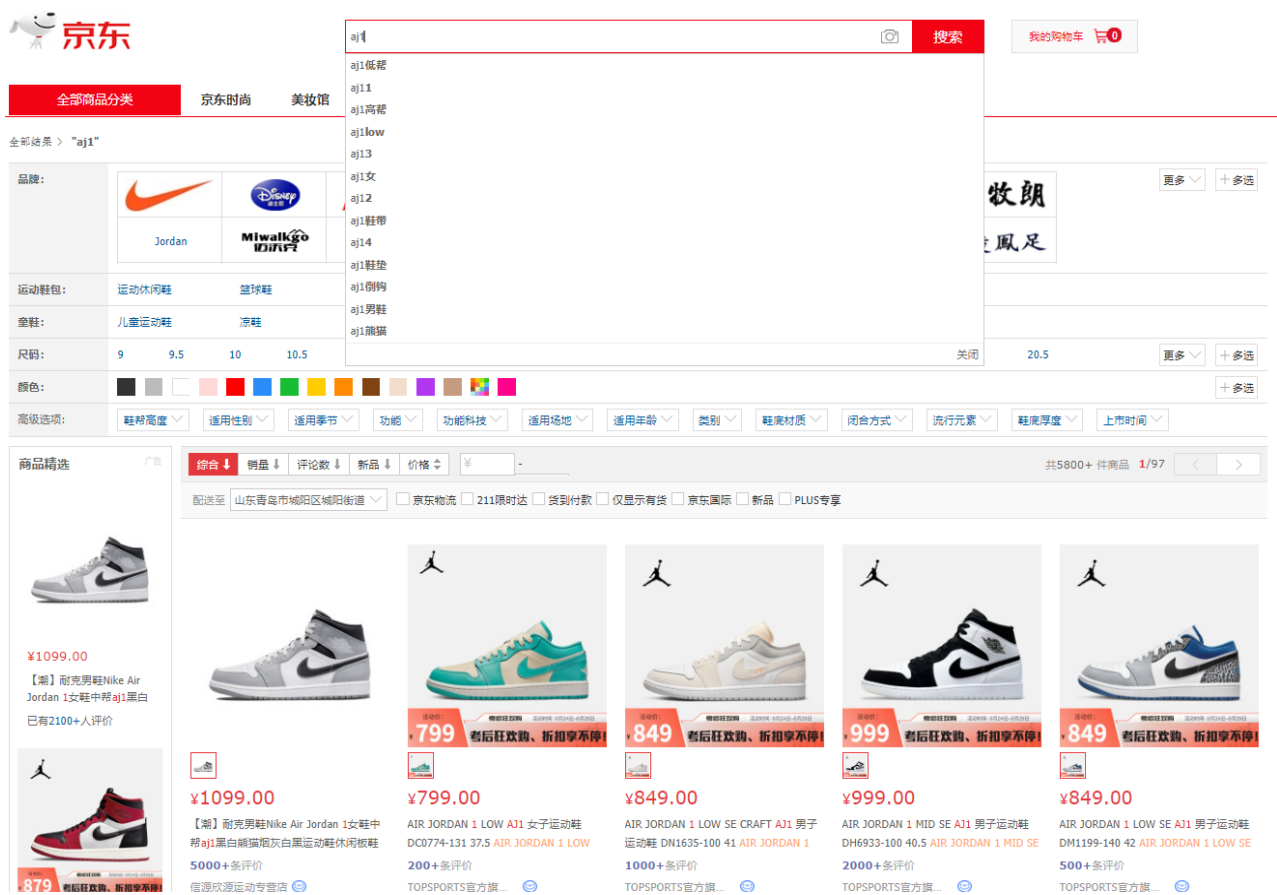


图 1 京东购物网站 ‘AJ1’ 搜索结果-商品展示

二、数据获取与清洗

1、数据描述

数据来源：京东商城[京东\(JD.COM\)-正品低价、品质保障、配送及时、轻松购物!](https://jd.com)

数据获取对象：‘AJ1 休闲鞋’商品的价格、售卖店铺、优先售卖鞋码等信息。

2、使用工具

python 是一种功能丰富的语言，它拥有一个强大的基本类库和数量众多的第三方扩展。报告中进行数据爬取使用到的库有：

①requests

②beautifulsoup

③ xlwt

3、数据获取步骤及代码

(1) 在京东购物网站上获取 ‘AJ1’ 搜索结果的 HTML 源码内容。

```
import requests
if __name__ == '__main__':
    url=https://search.jd.com/Search?keyword=AJ1&enc=utf-8' # 目标网站地址
    response = requests.get(url) # 发送请求访问网站
    html = response.text # 获取网页源码
    print(html) # 将源码打印到控制台
```

> <script>window.location.href='https://passport.jd.com/uc/1
== RESTART: C:/Users/yeahamen/AppData/Local/Programs/Pytho
Squeezed text (4519 lines).

(2) 在源码中筛选所需 ‘AJ1’ 商品相关信息数据

```
# 初始化 BeautifulSoup 库,并设置解析器
soup = BeautifulSoup(html, 'lxml')
goods_list = soup.find_all('li', class_='gl-item') # 在源码中获取商品列表

for li in goods_list: # 遍历商品列表获取下面信息
    no = li['data-sku'] # 商品编号
    name = li.find(class_='p-name p-name-type-2').find('em').get_text() # 商品名称
    price = li.find(class_='p-price').find('i').get_text() # 价格
    shop = li.find(class_='p-shop').find('a').get_text() # 商家
    detail_addr = li.find(class_='p-name p-name-type-2').find('a')['href'] # 商  
品详情地址
```

```

===== RESTART: C:/Users/yeahamen/Desktop/论文/获取AJ1数据.py =====
=====
--> 正在获取网站信息
=====分割线=====
商品编号 = 10044334096764
商品名称 = AIR JORDAN 1 ZOOM AIR CMFT AJ1 女子运动鞋 CT0979-610 36.5
价格 = 879.00
商家 = TOPSPORTS官方旗舰店
商品详情地址 = //item.jd.com/10044334096764.html
=====分割线=====
商品编号 = 10052975025718
商品名称 = AIR JORDAN 1 LOW AJ1 女子运动鞋 DC0774-131 37.5
价格 = 799.00
商家 = TOPSPORTS官方旗舰店
商品详情地址 = //item.jd.com/10052975025718.html
=====分割线=====

```

(3) 将获得的商品数据写入 Excel 文件。

```

# 创建 workbook，就是创建一个 Excel 文档

write_work = xlwt.Workbook(encoding='ascii') # 添加一张单

write_sheet = write_work.add_sheet("sheet1") # 创建表头

write_sheet.write(0, 0, label='商品编号')      # 第 1 行 第 1 列 写入内容'商品编号'
write_sheet.write(0, 1, label='商品名称')      # 第 1 行 第 2 列 写入内容'商品名称'
write_sheet.write(0, 3, label='价格')          # 第 1 行 第 3 列 写入内容'价格'
write_sheet.write(0, 4, label='商家')          # 第 1 行 第 4 列 写入内容'商家'
write_sheet.write(0, 5, label='商品详情地址')  # 第 1 行 第 5 列 写入内容'商品详情地址'

```

	A	B	C	D	E
1	商品编号	商品名称	价格	商家	商品详情地址
2	10026215122848	【HOT】耐克男鞋 Nike Air Jordan Mid aj1 熊猫 中帮运动休闲透气轻便篮球鞋 554725/554724-132 / 纯黑/黑绿 42.5	1049	考锐运动专营店	//item.jd.com/10026215122848.html
3	10022265004442	【潮】Nike Air Jordan aj1 Mid 耐克男鞋春季中帮透气休闲板鞋轻便运动篮球鞋 852542-101 40	1019.00	加百利运动专营店	//item.jd.com/10022265004442.html
4	55711792494	【潮】耐克男鞋 Nike Air Jordan 1 女鞋中帮 aj1 黑白熊猫等带绿 运动鞋休闲板鞋 篮球鞋 554725/554724-411 (黑曜石) 36.5	1339.00	信源欣源运动专营店	//item.jd.com/55711792494.html
5	10026408691861	【尖货】Nike Air Jordan aj1 耐克板鞋男女春季情侣新款中帮透气休闲轻便运动篮球鞋 DJ4695/554724-122 (白黑红) 42	1169.00	大生 (深圳) 运动户外专营店	//item.jd.com/10026408691861.html
6	10030789099605	【好物推荐】nike air jordan 1 Low 耐克男鞋春季新款 aj1 低帮板鞋运动篮球鞋 553558-612/553560-612 黑红脚趾 41	1359.00	领街运动户外专营店	//item.jd.com/10030789099605.html
7	10043218282336	AIR JORDAN 1 ZOOM AIR CMFT AJ1 女子运动鞋 DQ5092 DQ5092-651 39	1199.00	TOPSPORTS官方旗舰店	//item.jd.com/10043218282336.html
8	10048455186942	NIKE 耐克板鞋 AIR Jordan AJ1 男子新款休闲运动篮球鞋 CT0978-016 CT0978-016 42	679.00	源恒运动专营店	//item.jd.com/10048455186942.html
9	100014127929	耐克 NIKE 女子 篮球鞋 气垫 乔1 AJ1 ZOOM AIR CMFT SE 运动鞋 CZ1360-401 亮蓝色 37.5 码	1199.00	耐克 (NIKE) 京东自营专区	//item.jd.com/100014127929.html
10	62272129134	【潮】Nike Air Jordan aj1 耐克板鞋男女夏季情侣新款中帮透气轻便休闲运动篮球鞋 DN4321/DH6933-100 (黑白熊猫) 40 (男)	1099.00	考利运动专营店	//item.jd.com/62272129134.html

4、完整程序

```

import requests

from bs4 import BeautifulSoup

import xlwt

def get_html(url):
    # 模拟浏览器访问

    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/101.0.4951.64 Safari/537.36 Edg/101.0.1210.53'
    }

    print("--> 正在获取网站信息")

    response = requests.get(url, headers=headers) # 请求访问网站

    if response.status_code == 200:

        html = response.text # 获取网页源码

        return html # 返回网页源码

    else:

        print("获取网站信息失败！")

```

```

if __name__ == '__main__':
    # 创建 workbook, 就是创建一个 Excel 文档
    write_work = xlwt.Workbook(encoding='ascii')

    # 添加一张单
    write_sheet = write_work.add_sheet("sheet1")

    # 创建表头
    write_sheet.write(0, 0, label='商品编号')      # 第 1 行 第 1 列 写入内容'商品编号'
    write_sheet.write(0, 1, label='商品名称')      # 第 1 行 第 2 列 写入内容'商品名称'
    write_sheet.write(0, 2, label='价格')          # 第 1 行 第 4 列 写入内容'价格'
    write_sheet.write(0, 3, label='商家')          # 第 1 行 第 5 列 写入内容'商家'
    write_sheet.write(0, 4, label='商品详情地址')  # 第 1 行 第 6 列 写入内容'商品详情地址'

    # 记录当前行数
    _current_row = 0

    i=0
    k=0
    j=0
    for k in range(0,5):
        i = 3+k*2
        j = 56+k*60
        # 搜索关键字
        keyword = 'AJ1'
        # 搜索地址
        search_url=
'https://search.jd.com/Search?keyword=%s&suggest=1.his.0.0&wq=AJ1&pvid=65892364c2604d5
897754ab21bed6d22&page=%d&s=%d&click=1'%(keyword,i,j)

        html = get_html(search_url)
        # 初始化 BeautifulSoup 库, 并设置解析器
        soup = BeautifulSoup(html, 'lxml')

        # 商品列表
        goods_list = soup.find_all('li', class_='gl-item')
        # 打印 goods_list 到控制台
        for li in goods_list: # 遍历父节点
            # 由于我们第一行已经写入了表头。所以这里 0+1, 就是从第 1 行开始, 后面每次循环+1
            _current_row += 1
            if _current_row == 29:
                break
            # 商品编号
            no = li['data-sku']
            # 商品名称
            name = li.find(class_='p-name p-name-type-2').find('em').get_text()
            # 价格
            price = li.find(class_='p-price').find('i').get_text()
            # 商家
            shop = li.find(class_='p-shop').find('a').get_text()
            # 商品详情地址
            detail_addr = li.find(class_='p-name p-name-type-2').find('a')['href']

```

```
# 写入 Excel

write_sheet.write(_current_row, 0, label=no)

write_sheet.write(_current_row, 1, label=name)

write_sheet.write(_current_row, 2, label=price)

write_sheet.write(_current_row, 3, label=shop)

write_sheet.write(_current_row, 4, label=detail_addr)

# 保存文件，使用的是相对目录（也可以使用绝对路径），会保存在当前文件的同目录下。文件名为读取多个商品页面 1.xls，必须是.xls 后缀

write_work.save("../读取多个商品页面 1.xls")
```

5 数据保存到 Excel 文件【读取多个商品页面 1.xls】，Excel 文件截图如图 2 所示：

商品编号	商品名称	价格	商家	商品详情地址
1004956542805	2022年四季新款AJ1高帮休闲潮流百搭板鞋空军一号篮球鞋男女透气运动鞋男鞋女鞋BBCANKS 极光蓝 43	¥228.00	BBCANKS旗舰店	/item.jd.com/1004956542805.html
10040942766362	AIR JORDAN 1 MID SE AJ1 男子运动鞋 DC7294-200 45	¥999.00	TOPSPORTS官方旗舰店	/item.jd.com/10040942766362.html
10026215122864	【HOT】耐克男鞋 Nike Air Jordan Mid aj1低帮 中帮运动休闲透气轻便篮球鞋 DJ4695/554724-122 白黑红 42	¥149.00	考锐运动专营店	/item.jd.com/10026215122864.html
10041329918499	AIR JORDAN 1 MID AJ1 男子运动鞋 554724-082 42	¥999.00	TOPSPORTS官方旗舰店	/item.jd.com/10041329918499.html
10049738442304	AIR JORDAN 1 LOW (GS) AJ1 运动鞋 DM8960-801 38	¥64.00	TOPSPORTS官方旗舰店	/item.jd.com/10049738442304.html
10023901717210	潮 Nike男鞋 Air Jordan aj1 low耐克低帮板鞋球鞋水洗黑中黑蓝UNC彩色拼接 DC6991-200 42	¥439.00	德鲁运动专营店	/item.jd.com/10023901717210.html
10052440346853	高帮联名HISERND男鞋球鞋鞋子女1高帮变色龙黑红男+南白黑脚趾侧边女鞋板鞋板鞋 AJ1高帮-倒勾 41	¥258.00	德鲁运动户外旗舰店	/item.jd.com/10052440346853.html
10030638589550	【好物推荐】耐克男鞋Nike Air Jordan aj1 low秋季新款板鞋低帮防侧滑运动耐穿篮球鞋 CK3022-301 黑绿钻石 40.5	¥249.00	德鲁运动户外专营店	/item.jd.com/10030638589550.html
10047761244834	AIR JORDAN 1 LOW ALT (PS) AJ1 运动鞋 DM8948-100 35	¥449.00	TOPSPORTS官方旗舰店	/item.jd.com/10047761244834.html
157680301122	【潮】Nike Air Jordan aj1 GS耐克板鞋女子夏季中帮透气休闲轻便运动耐穿篮球鞋 CT0979-101 (彩色拼接 马卡龙) 37.5	¥299.00	考锐运动专营店	/item.jd.com/157680301122.html
10052895243007	Nike耐克 Air Jordan 1 LOW AJ1 高帮 男女低帮板鞋古铜球鞋 DV1762-001 43	¥1009.00	北京德惠运动专营店	/item.jd.com/10052895243007.html
10041370722650	【尖货】Nike Air Jordan aj1耐克板鞋男女春季情侣新款中帮透气休闲轻便耐穿运动篮球鞋 554725/554724-411 (黑曜石) 39	¥1059.00	大生 (深圳) 运动户外专营店	/item.jd.com/10041370722650.html
10042694511613	2022年春季新款aj1高帮板鞋aj1春季新款变色龙男女鞋BBCANKS 灰蓝 43	¥188.00	BBCANKS旗舰店	/item.jd.com/10042694511613.html
10042694511613	AIR JORDAN 1 ZOOM AIR CMFT AJ1 男子运动鞋 DQ5091-041 42	¥1079.00	TOPSPORTS官方旗舰店	/item.jd.com/10042694511613.html
11764264487	【潮】耐克男鞋Nike Air Jordan 1(鞋中帮aj1)黑白绿黑灰白黑运动鞋休闲板鞋 篮球鞋 554724-078 41	¥1249.00	德鲁运动专营店	/item.jd.com/11764264487.html
153769511439	AIR JORDAN 1 RET LOW SLIP AJ1 女子运动鞋 AV3918 AV3918-600 36	¥439.00	TOPSPORTS官方旗舰店	/item.jd.com/153769511439.html
10052085574142	NIKE耐克 Air Jordan 1 LOW SE AJ1 复古篮球鞋 DN1635-100 42	¥839.00	北京德惠运动专营店	/item.jd.com/10052085574142.html
18170789845	【潮】Nike耐克男鞋 Air Jordan 1 Low AJ1 男1黑脚趾低帮板鞋休闲运动复古篮球鞋 553558-123 (白蓝黄) 40	¥1159.00	君悦运动专营店	/item.jd.com/18170789845.html
10052549123599	秋季高帮男1金篮球鞋AJ1高帮男子黑脚趾高帮板鞋运动耐穿公司级南白鞋 白红 44	¥418.00	德惠运动户外旗舰店	/item.jd.com/10052549123599.html
10044584948136	AIR JORDAN 1 MID (GS) AJ1 运动鞋 554725-604 37.5	¥799.00	TOPSPORTS官方旗舰店	/item.jd.com/10044584948136.html
10052006673935	ysports Nike耐克 Air Jordan 1 Mid AJ1 女子高帮篮球鞋BQ8472 BQ6472-161 37.5	¥999.00	Yysports旗舰店	/item.jd.com/10052006673935.html
10026410007980	【尖货】Nike Air Jordan aj1耐克板鞋女子春季新款低帮透气休闲耐穿耐穿运动篮球鞋 554723-106 38	¥1009.00	大生 (深圳) 运动户外专营店	/item.jd.com/10026410007980.html
10045314166312	AIR JORDAN 1 MID SE (GS) AJ1 运动鞋 DM1208-150 35.5	¥699.00	TOPSPORTS官方旗舰店	/item.jd.com/10045314166312.html
10048055947760	NIKE耐克 Air Jordan 1 Zoom Air AJ1 男1高帮 男女复古篮球鞋 DQ0659-700(男女同款) 35.5	¥609.00	北京德惠运动专营店	/item.jd.com/10048055947760.html
10044342273047	AIR JORDAN 1 ELEMENT AJ1 男子运动鞋 DB2889-100 42	¥499.00	TOPSPORTS官方旗舰店	/item.jd.com/10044342273047.html
10039411367360	AIR JORDAN 1 LOW UTL (GS) AJ1 运动鞋 DQ2233 DQ2233-264 37.5	¥499.00	TOPSPORTS官方旗舰店	/item.jd.com/10039411367360.html
10038894132868	AIR JORDAN 1 MID SE AJ1 男子运动鞋 DN4904-001 44	¥999.00	TOPSPORTS官方旗舰店	/item.jd.com/10038894132868.html
10024904942251	烽火 Air Jordan 1 AJ1 男1 小钢炮TS 黑摩卡 黑绿 篮球鞋 555088 105 555088-105 12号XB会现货 44	¥4049.00	烽火体育SNEAKER之家	/item.jd.com/10024904942251.html
10051808576203	【正货】高帮低帮1/1 艾世纯原 Air AJ35 "Dynasties" 紫禁城制霸实战篮球鞋 紫禁 39	¥26.00	流江运动服饰旗舰店	/item.jd.com/10051808576203.html
10034661767912	Nike耐克 Air Jordan 1 LOW SE 低帮运动休闲篮球鞋板鞋 553558-118/553560-118 芝加哥 36.5	¥1159.00	德惠运动户外专营店	/item.jd.com/10034661767912.html
153558623708	【潮】Nike Air Jordan aj1 Mid耐克板鞋女子春季新款中帮透气休闲轻便运动篮球鞋 DH0210-100 黑杏粉 36.5	¥989.00	加加运动专营店	/item.jd.com/153558623708.html
10049847325796	ysports Nike耐克 Air Jordan 1 Mid AJ1 中帮篮球鞋BQ8472 DQ6699-200 36	¥999.00	Yysports旗舰店	/item.jd.com/10049847325796.html
59091318993	【潮】Nike Air Jordan aj1耐克板鞋男女春季新款中帮透气休闲耐穿耐穿运动篮球鞋 554725/554724-123 薄荷绿 42.5	¥1069.00	考锐运动专营店	/item.jd.com/59091318993.html
10049172001164	nike耐克男鞋AIR JORDAN 1运动篮球鞋CD7069-118 DB2889-700 [AJ1] 44.5	¥69.00	黑石旗舰店	/item.jd.com/10049172001164.html
10053718993690	【官方自营旗舰店】aj男女鞋莆田新款字母哥3支球队团队 德惠实战耐穿训练篮球鞋GT8(CDH4528-100 字母哥 玄绿 蓝白 莆田纯原版本 43 极	¥417.00	华联来运动户外旗舰店	/item.jd.com/10053718993690.html
10033119177397	AIR JORDAN 1 MID SE (GS) AJ1 运动鞋 DC4099-100 37.5	¥689.00	TOPSPORTS官方旗舰店	/item.jd.com/10033119177397.html

图 2 ‘AJ1’ 休闲鞋商品价格等数据的 excel 文件截图

三、数据处理和分析

1 数据可视化工具

python 是一种功能丰富的语言，它拥有一个强大的基本类库和数量众多的第三方扩展。报告中进行分析与可视化所用到的库有：

- ①pandas
- ②matplotlib
- ③Counter
- ④openpyxl
- ⑤numpy

注：可以带着这样的目的进行本次数据分析：有一位男性朋友鞋码 42，拿着 1000 元想买一双 AJ1，如何给他做出推荐？

2 数据分析与可视化（鞋码分析）

(1) 从原始文件中提取‘鞋码’数据至单独 Excel 文件。

```
xiema = [] #建立空列表
excel=xlrd.open_workbook("读取多个商品页面.xls") #打开 excel 文件
sheet=excel.sheet_by_index(0) #根据下标获取工作簿，这里获取第一个
good = sheet.col_values(1) #获取第二列的内容
#遍历每一个商品名称，下面 sss 指的是一些字符
for i in range(len(good)):
    #匹配以数字+数字+除\n 以外任意字符+数字结尾的数字
    if re.findall('\d\d\.\d$', good[i]) != []:
        xiema.append(re.findall('\d\d\.\d$', good[i]))
    #匹配以实数结尾的数字如：sssss42
    elif re.findall('\d\d$', good[i]) != []:
        xiema.append(re.findall('\d\d$', good[i]))
    #匹配（数字+数字+除\n 以外任意字符+数字）+码，但只保留括号内的内容
    elif re.findall('(\d\d\.\d)码', good[i]) != []:
        xiema.append(re.findall('(\d\d\.\d)码', good[i]))
    #匹配空白字符/制表符/回车+数字+数字
    elif re.findall('\s\d\d', good[i]) != []:
        #匹配时+空白符，添加到列表时只添加数字
        xiema.append(re.findall('\s(\d\d)', good[i]))
print(xiema) #【下面是打印出来的鞋码列表】
```

```
[['42.5'], ['40'], ['36.5'], ['42'], ['41'], ['39'], ['42'], ['37.5'], ['40'], ['43'], ['43'], ['35'], ['38'], ['36'], ['39'], ['40.5'], ['36'], ['36'], ['42'], ['37.5'], ['27'], ['42'], ['39'], ['38'], ['36'], ['37.5'], ['42'], ['36'], ['27'], ['43'], ['43'], ['41'], ['37.5'], ['45'], ['36'], ['40.5'], ['38.5'], ['35.5'], ['40.5'], ['39'], ['37.5'], ['44.5'], ['38'], ['40'], ['42'], ['42'], ['37.5'], ['41'], ['42.5'], ['41'], ['42'], ['41'], ['42.5'], ['40'], ['40.5'], ['41'], ['42.5'], ['40'], ['40'], ['40'], ['40'], ['37.5'], ['38.5'], ['38.5'], ['36'], ['37.5'], ['41'], ['40.5'], ['43'], ['38.5'], ['36'], ['43'], ['40'], ['40'], ['40.5'], ['40.5'], ['41'], ['38'], ['42.5'], ['42.5'], ['40'], ['42'], ['36'], ['38'], ['39'], ['38.5'], ['40.5'], ['38.5'], ['42'], ['38.5'], ['42.5'], ['41'], ['44.5'], ['44'], ['40'], ['43'], ['44.5'], ['42.5'], ['36.5'], ['43'], ['38'], ['35.5'], ['36.5'], ['43'], ['44'], ['40'], ['36'], ['40'], ['37.5'], ['42'], ['38.5'], ['41'], ['43'], ['41'], ['41'], ['40'], ['36.5'], ['44'], ['40'], ['36'], ['38'], ['37.5'], ['40'], ['42'], ['39'], ['39'], ['38'], ['42'], ['44'], ['38'], ['38'], ['42.5'], ['40'], ['42.5'], ['42.5'], ['40'], ['28'], ['35.5'], ['37.5'], ['44'], ['38'], ['37.5'], ['36'], ['36'], ['40.5'], ['37.5'], ['36'], ['36.5']]
```

```
write_work = xlwt.Workbook(encoding='ascii') # 创建 workbook，就是创建一个 Excel 文档
write_sheet = write_work.add_sheet("sheet1") # 添加一张表单
for i in range(0, len(xiema)):
    write_sheet.write(i, 0, xiema[i]) # 第 i 行 第 1 列 写入内容鞋码
write_work.save("./鞋码写入到 excel 里面.xls") # 下面是把鞋码存入单独的文件
```


鞋码数据从原始文件中提取到 Excel 文件【鞋码写入到 excel 里面.xls】

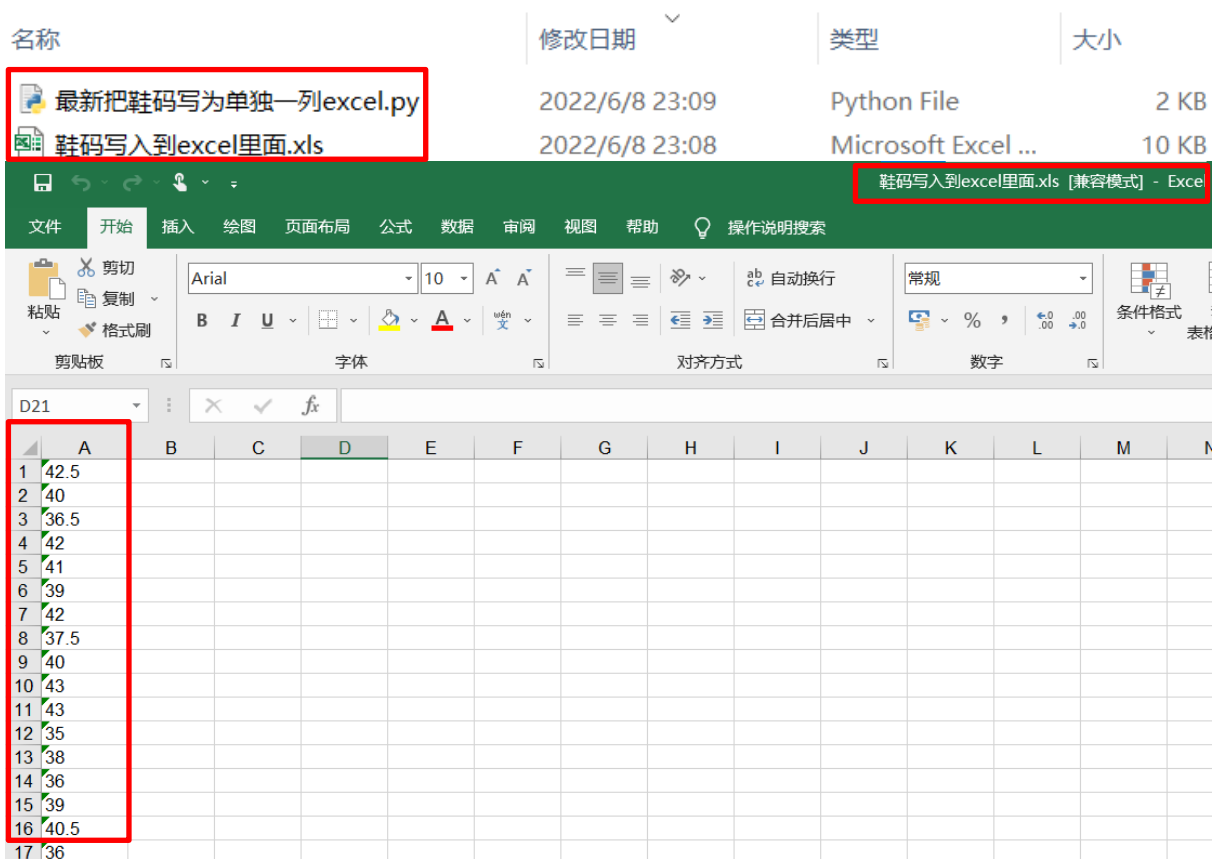


图 3 ‘AJ1’ 休闲鞋 鞋码数据的 excel 文件截图

(2) 统计所有鞋码分布范围

```
data = [] #建立空列表
for i in xiema: #遍历刚才写好的列表
    data.append(float(i[0])) #提取每一个鞋码浮点化
less39 = 0 #定义三个变量接受三种频数
morethan42 = 0
middle = 0
for i in data: #确定范围内频数
    if i < 39:
        less39 += 1
    elif i > 42:
        morethan42 += 1
    else:
        middle += 1
biaotou = ['小于 39 号鞋码数', '39-42 号鞋码数', '大于 42 号鞋码数']
# 创建 workbook, 就是创建一个 Excel 文档
write_work = xlwt.Workbook(encoding='ascii')
write_sheet = write_work.add_sheet("sheet1") # 添加一张表单
for i in range(0,3): # 创建表头
    write_sheet.write(i,0,biaotou[i]) # 第 i 行 第 1 列 写入内容鞋码
```

```

write_sheet.write(0,1,less39)
write_sheet.write(1,1,middle)
write_sheet.write(2,1,morethan42)
write_work.save("./鞋码的分布情况 excel 文件.xls")

```

将鞋码分布范围写入 Excel 文件【鞋码的分布情况 excel 文件.xls】

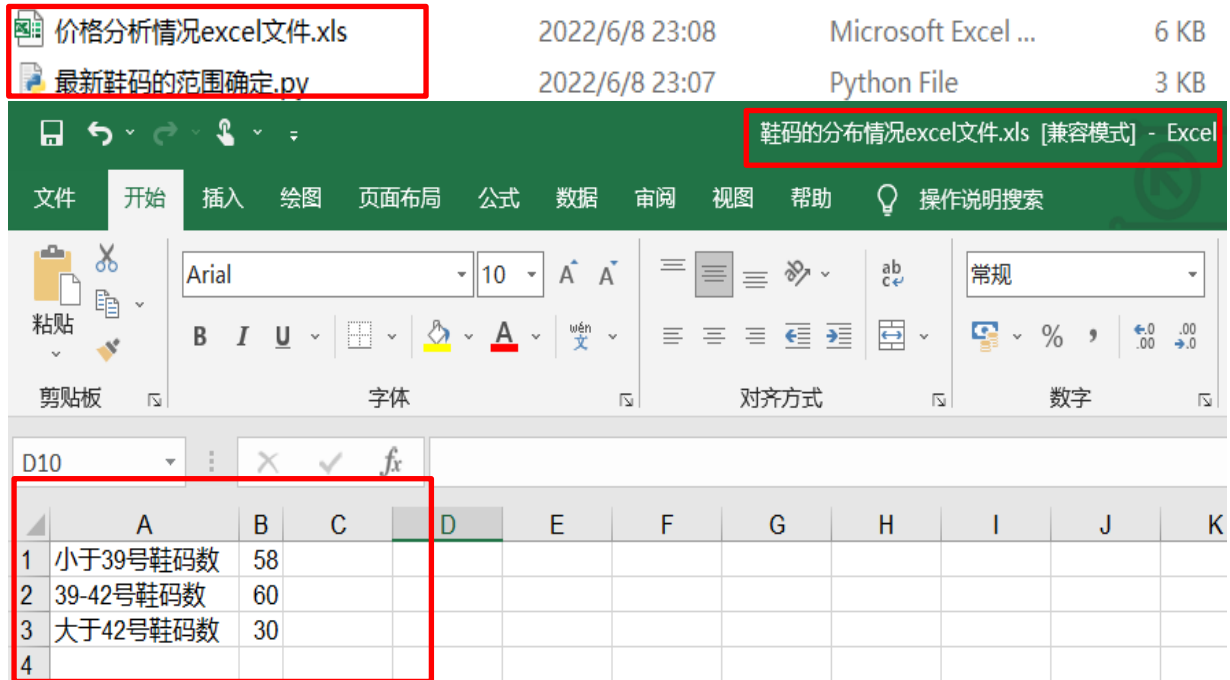


图4 ‘AJ1’ 休闲鞋 鞋码数据分布范围的 excel 文件截图

(3) 鞋码分布范围的可视化【饼状图】

```

import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import rcParams
plt.rcParams['font.sans-serif']=['Microsoft YaHei']#显示中文标签,处理中文乱码问题
plt.rcParams['axes.unicode_minus']=False #坐标轴负号的处理
plt.axes(aspect='equal') #将横、纵坐标轴标准化处理,确保饼图是一个正圆,否则为椭圆
rcParams['font.family'] = 'simhei'
text=r"C:\Users\86178\Desktop\鞋码的分布情况 excel 文件.xls"#获取路径
pe=pd.read_excel(text) #读取文件
xx=pe['范围'] #获取数据
yy=pe['数量']
edu = yy #构造数据
labels = xx
colors = ['#9999ff', '#ff9999', '#7777aa', '#2442aa', '#dd5555'] #自定义颜色
plt.pie(x=edu, labels=labels, autopct='% .2f%',)
plt.title('鞋码分布饼状图',fontsize=30) #添加图标题,设置字大小
plt.show() #显示图形

```

绘制图像【鞋码分布饼状图】

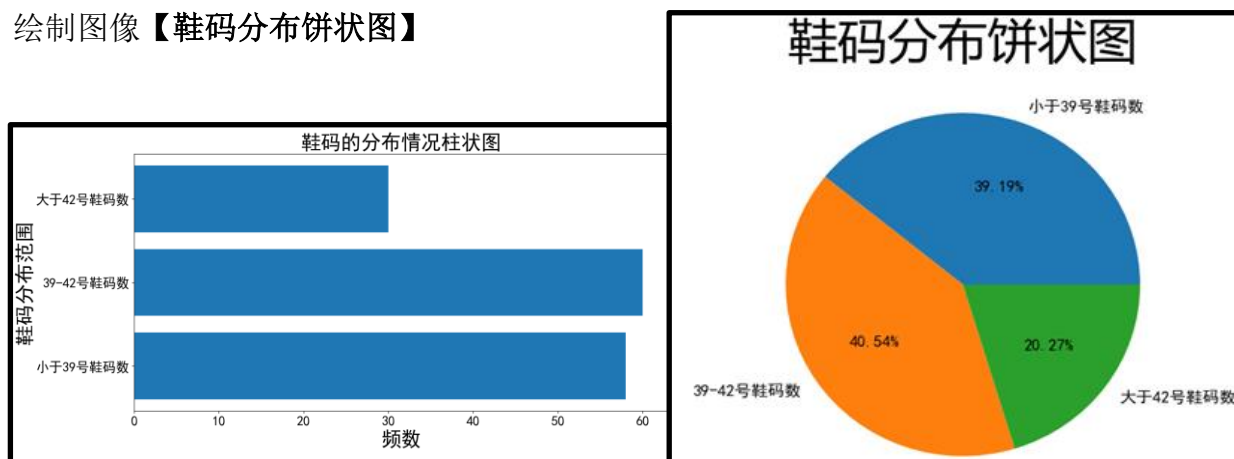


图 5 ‘AJ1’ 休闲鞋 鞋码数据分布范围的饼状图

小结:

根据调查: 中国成年男性的足长平均为 270cm 对应鞋码 41-42

中国成年女性的足长平均为 245cm 对应鞋码 38-39

对应图 5: 鞋码分布饼状图

- ① 39-42 号的鞋码可视为成年男性的穿鞋标准;
- ② <39 号的鞋码可以视为成年女性的穿鞋标准;
- ③ >42 号的鞋码可以视为身材比较高大人群的穿鞋标准。

前两者在大众中占比较高且相近, 后者占比较少是众所周知。

图 5 很明显地显示出: 网站地鞋码推荐比例与社会鞋码地分布有很好地相关性。

因此可以得出结论: AJ1 适合所有成年人购买穿戴。朋友的 42 号穿鞋标准可以购买 AJ1 鞋。

3 数据分析与可视化 (价格、商家分析)

(1) 从原始文件中提取 ‘价格’、‘商家’ 数据至单独 Excel 文件。

```
import pandas as pd #导入库
df = pd.read_excel(r'读取多个商品页面.xls', usecols=[2, 3]) #从原文件中读出第 3、4 列
df_li = df.values.tolist() #将每列数据转化成一个列表
df = pd.DataFrame(df_li, columns=['价格', '商家']) #建立 dataframe 类型
df.to_excel("价格-商家.xls", index=False) #写入文档 #写入新文档: 价格-商家 excel
```

写入 Excel 文件【价格—商家.xls】



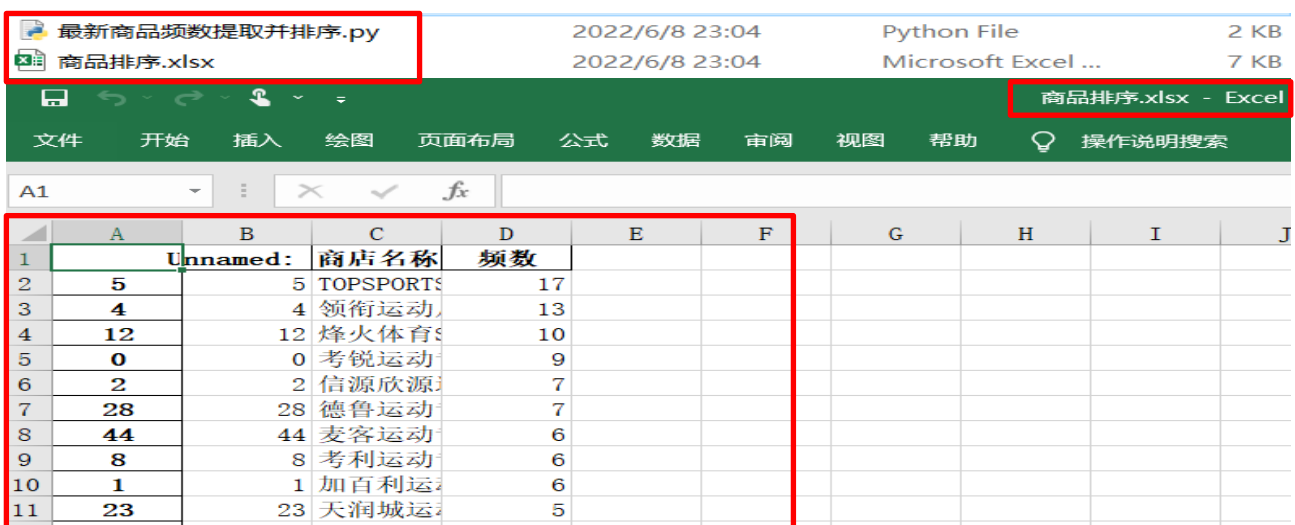
图 6 ‘AJ1’ 休闲鞋 价格-商家数据存储 Excel 文件截图

(2) 从文件【价格—商家.xls】中汇总每个商家出现的频数写入到单独 Excel 文件。

```
from collections import Counter
import pandas as pd

data = pd.read_excel('价格-商家.xlsx')      # 导入价格-商家信息
x_pandas_list = data[u'商家']               # 获取“商家”栏目数据
list1 = list(x_pandas_list)                 # 转化列表类型
c = Counter(list1)                         # 自动提取相同店铺次数并排序(升)
d = dict(c)                                # 转为字典格式(顺序乱)
print(c)                                   # 输出
key = list(d.keys())                       # 提取字典 d 中的所有键
value = list(d.values())                   # 提取 d 中所有的“值”
result_excel = pd.DataFrame()              # 利用 pandas 模块先建立 DataFrame 类型
result_excel["商店名称"] = key             # 把 key、value 列表存进去
result_excel["频数"] = value
result_excel.to_excel('商品频率.xlsx')     # 写入 excel
data= pd.read_excel('商品频率.xlsx')       # 读取前面写好的键-值 excel
data = data.sort_values(by = ['频数'],ascending = False)#排序(升)
print(data)                               # 输出
data.to_excel('商品排序.xlsx')             # 写入文件(排序完成)
```

写入 Excel 文件【商品排序.xlsx】(同时也生成了没有进行排序的基本文件商品频率.xlsx)



	A	B	C	D	E	F	G	H	I	J
1	5	5 TOPSPORTS	17							
2	4	4 领街运动	13							
3	12	12 烽火体育	10							
4	0	0 考锐运动	9							
5	2	2 信源欣源	7							
6	28	28 德鲁运动	7							
7	44	44 麦客运动	6							
8	8	8 考利运动	6							
9	1	1 加百利运	6							
10	23	23 天润城运	5							

图 7 ‘AJ1’ 休闲鞋 商家频数排序存储 Excel 文件截图

(3) 从文件【价格—商家.xls】中汇总每个商家的平均售价写入到单独 Excel 文件。

```
import pandas as pd
import openpyxl

wb=openpyxl.load_workbook(r'商品频率.xlsx') #读取文件商品频率
ws=wb.active                                #获取工作表索引值,无更改则一直工作这一张
data = pd.read_excel(r'读取多个商品页面.xls') #读取文件‘读取多个商品页面’给 data
data1 = pd.read_excel(r'商品频率.xlsx')       #读取文件‘商品频率’给 data1
name1=data1['商店名称']                     #读取 data1‘商店名称’列(50)
shangjia = data['商家']                     #读取 data‘商家’列(148)
jiage=data['价格']                          #读取 data‘价格’列(148)
```

```

name=[]
for i in shangjia:                #把 data`商家` 列给 name 列表 (148)
    if type(i)!=str:              #把空数据筛选出去 (读取网页时的空行)
        continue
    name.append(i)

name=list(set(name))              #set 是创建一个无序不重复的可迭代结构
print(name)                      #把商家名称转化为 `name` 列表 (50)
al_price=[]                      #创建下面要用到的新列表 `al_price`
for i in name:                   #遍历每一个不重复的商家 (50)
    price=0                      #定义两个变量来求平均值
    counter=0
    for j in range(len(shangjia)): #遍历每一个商家 (148)
        if i==shangjia[j]:      #求所有商家 i 的价格之和
            price+=jiage[j]
            counter+=1
    al_price.append(price/counter) #平均价格都加入到新列表 al_price
data2=zip(name,al_price)         #把不重复商家及其售价平均值合并到元组
for i in data2:                 #遍历元组
    for j in range(len(name1)):  #遍历商店名称
        if str(i[0])==str(name1[j]):
            ws.cell(row=j+2,column=4,value=i[1]) #把平均值写入商店名对应的行
    wb.save(r'商品频率.xlsx')    #保存文件

data= pd.read_excel('商品频率.xlsx') #读取前面写好的键-值 excel
data = data.sort_values(by = ['频数'],ascending = False)#排序 (升)
data.to_excel('商品频率.xlsx')    #写入文件 (排序完成)

```

把平均价格一并写入 Excel 文件【商品频率.xlsx】

名称	修改日期	类型	大小
读取多个商品页面.xls	2022/6/3 22:04	Microsoft Excel ...	71 KB
商品频率.xlsx	2022/6/24 23:03	Microsoft Excel ...	8 KB
最新获取商家频数及平均价格.py	2022/6/24 23:03	Python File	3 KB
最新排序生成商品频率.py	2022/6/24 23:01	Python File	2 KB

商品频率.xlsx - Excel					
文件	开始	插入	绘图	页面布局	公式
数据	审阅	视图	帮助	操作说明搜索	

	A	B	C	D	E	F	G
1		Unnamed: 0	商店名称	频数	平均价格		
2	5	5	TOPSPORTS官方旗舰店	17	833.7058824		
3	4	4	领街运动户外专营店	13	1849.769231		
4	12	12	烽火体育SNEAKER之家	10	1491		
5	0	0	考锐运动专营店	9	1335.666667		
6	2	2	信源欣源运动专营店	7	1474.714286		
7	28	28	德鲁运动专营店	7	1893.285714		
8	44	44	麦客运动专营店	6	1489		
9	8	8	考利运动专营店	6	1712.333333		
10	1	1	加百利运动专营店	6	1307.333333		
11	23	23	天润城运动户外专营店	5	1633.2		
12	29	29	阿塔克运动专营店	4	2234		
13	22	22	飒威运动专营店	4	1154		
14	15	15	北京博恩运动专营店	3	1132.333333		
15	35	35	bebe8运动专营店	3	1272.333333		

图8 ‘AJ1’ 休闲鞋 商家平均价格（售价）存储 Excel 文件截图

(4) 绘制平均价格-店铺名称-条方图

```
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import rcParams

plt.rcParams['font.sans-serif']=['Simhei'] #设置中文字体
plt.rcParams['axes.unicode_minus']=False # 显示符号
dt=pe=pd.read_excel(r"商店频率.xls") #读取文件
pe.plot.bar(x='商店名称',y='平均价格') #绘制柱状图
plt.show() #展示图像
```

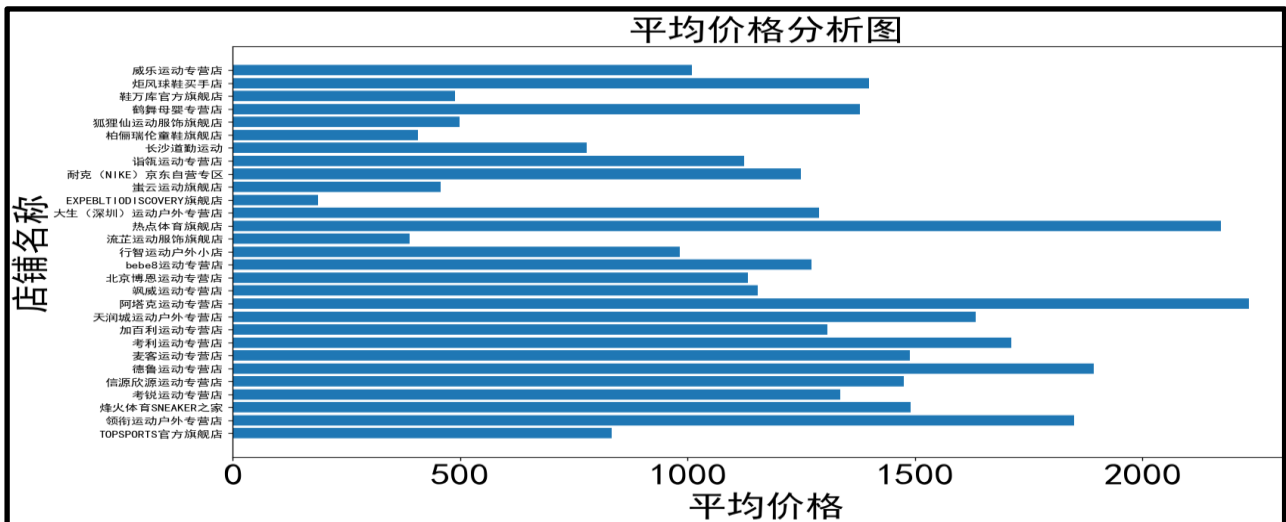


图9 ‘AJ1’ 休闲鞋 商家平均价格（售价）分析条方图

小结：

此条方图为消费者提供直观的价格展示，可根据自己的价格需求直接选择店铺购买；可以推荐拥有 1000 元朋友去 TOPSPORTS 官方旗舰店、行智运动户外小店、长沙道勤运动等店铺购买 AJ1。

(5) 把平均价格的范围分布汇总到单独 Excel 文件。

```
#价格统计
import xlrd
import xlwt
import csv
import re
import pandas as pd



xiema = [] #建立空列表
excel=xlrd.open_workbook("读取多个商品页面1.xls") #打开 excel 文件
sheet=excel.sheet_by_index(0) #根据下标获取工作簿，这里获取第一个
jiage = sheet.col_values(2) #获取第二列的内容
#价格统计
range1 = 0
range2 = 0
range3 = 0
range4 = 0
```

```

range5 = 0
for i in range(0,len(jiage)):
    if jiage[i] == '' or jiage[i] == '价格': #如果为空就要给变量赋默认值 5
        jiage[i] = 0
    else:
        jiage[i] = float(jiage[i])
for i in jiage:
    if i == '价格':
        continue
    if int(i) >= 2000:
        range1 += 1
    elif 1500 < int(i) < 2000:
        range2 += 1
    elif 1000 < int(i) < 1500:
        range3 += 1
    elif 500 < int(i) < 1000:
        range4 += 1
    else:
        range5 += 1
biaotoul=['>=2000','2000-1500','1500-1000','1000-500','0-500']
# 创建 workbook, 就是创建一个 Excel 文档
write_work = xlwt.Workbook(encoding='ascii')
# 添加一张单
write_sheet = write_work.add_sheet("sheet1")
# 创建表头
for i in range (0,5):
    write_sheet.write(i,0,biaotoul[i])      # 第 i 行 第 1 列 写入内容鞋码
write_sheet.write(0,1,range1)
write_sheet.write(1,1,range2)
write_sheet.write(2,1,range3)
write_sheet.write(3,1,range4)
write_sheet.write(4,1,range5)
write_work.save("./价格范围分布情况 excel 文件.xls")

```

写入 Excel 文件【价格范围分布情况 excel 文件.xls】

	价格范围分布情况excel文件.xls	2022/6/24 23:26	Microsoft Excel ...	6 KB
	价格分布范围分析.py	2022/6/24 23:26	Python File	2 KB

	A	B	C	D	E	F	G	H	I	J	K
1	>=2000	14									
2	2000-1500	13									
3	1500-1000	45									
4	1000-500	50									
5	0-500	28									

图 10 ‘AJ1’ 休闲鞋 商家平均价格分布范围文件截图

(6) 绘制平均价格范围分布雷达图

```
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import rcParams
import numpy as np

matplotlib.rcParams['font.family']='SimHei' #将字体设置为黑体'SimHei'
matplotlib.rcParams['font.sans-serif'] = ['SimHei']
rcParams['font.family'] = 'simhei' #上面是导入库和画图准备
pe=pd.read_excel(r"C:\Users\86178\Desktop\价格分析情况 excel 文件.xls")
x=pe['范围'] #获取图表数据
y=pe['频数']
labels =x
data = y

dataLenth = 5 #雷达分区数
angles = np.linspace(0,2*np.pi,dataLenth,endpoint=False) #0 到 2π 角度上找五个位置,
data = np.concatenate((data,[data[0]])) #使数据首尾相接, 使图形闭合
angles = np.concatenate((angles,[angles[0]]))
labels=np.concatenate((labels,[labels[0]]))

fig = plt.figure(facecolor="white") #绘图
plt.subplot(111,polar=True) #设置图形为极坐标图, 一个角度对应一个
值

plt.plot(angles,data,'bo-',color='g',linewidth=2) #连接上行代码的数值点, 画折线图 green
plt.fill(angles,data,facecolor='g',alpha=0.25) #填充两条线之间的色彩, alpha 为透
明度

plt.thetagrids(angles*180/np.pi,labels) #画出角度刻度并添加标签
plt.figtext(0.52,0.95,'平均价格范围',fontsize=20,ha='center') #添加雷达图标题
plt.grid(True)
plt.show()
```

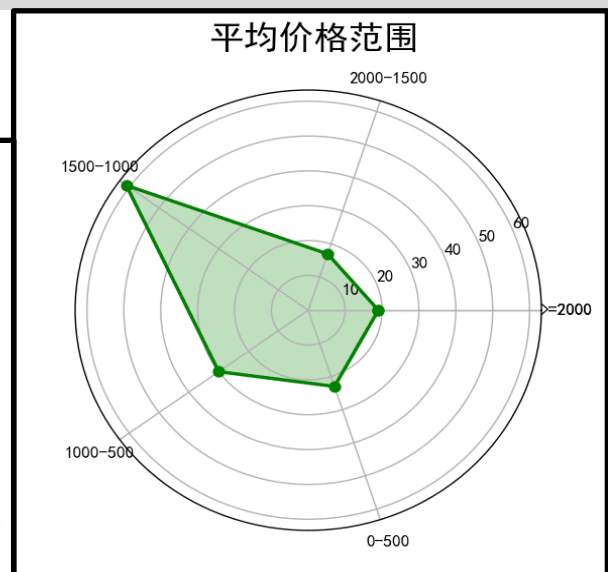
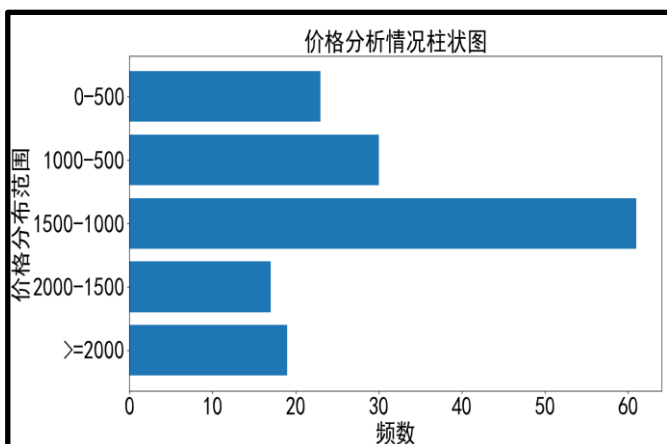


图 11 ‘AJ1’ 休闲鞋 商家平均价格分布范围雷达图

小结:

观察雷达图不难发现, 京东平台的 AJ1 鞋平均价格在 1000-1500 的最多;

朋友的 1000 元可能会偏少, 建议他多准备 100 或 200 元。

(7) 绘制商家频数分布条方图

```
import pandas as pd

import matplotlib.pyplot as plt
from matplotlib import rcParams

plt.rcParams['font.sans-serif']=['Simhei'] #设置中文字体
plt.rcParams['axes.unicode_minus']=False# 显示符号

pe=pd.read_excel(r"商品频率.xls")      #读取文件
k=pe['商店名称']                        #获取数据列
m=pe['频数']

plt.xlabel('商家频数',fontsize=30)      #横轴名称
plt.ylabel('商家名称',fontsize=30)      #纵轴名称

plt.xticks(fontsize=30)                 #调整横坐标数字大小
plt.title('商家频数分析图', fontsize=30)#设置图表名称及大小

plt.barh(k,m)                           #画图
plt.show()                              #展示
```

并且将商家频数条方图与前面所绘制商家平均价格条方图放在一起对比:

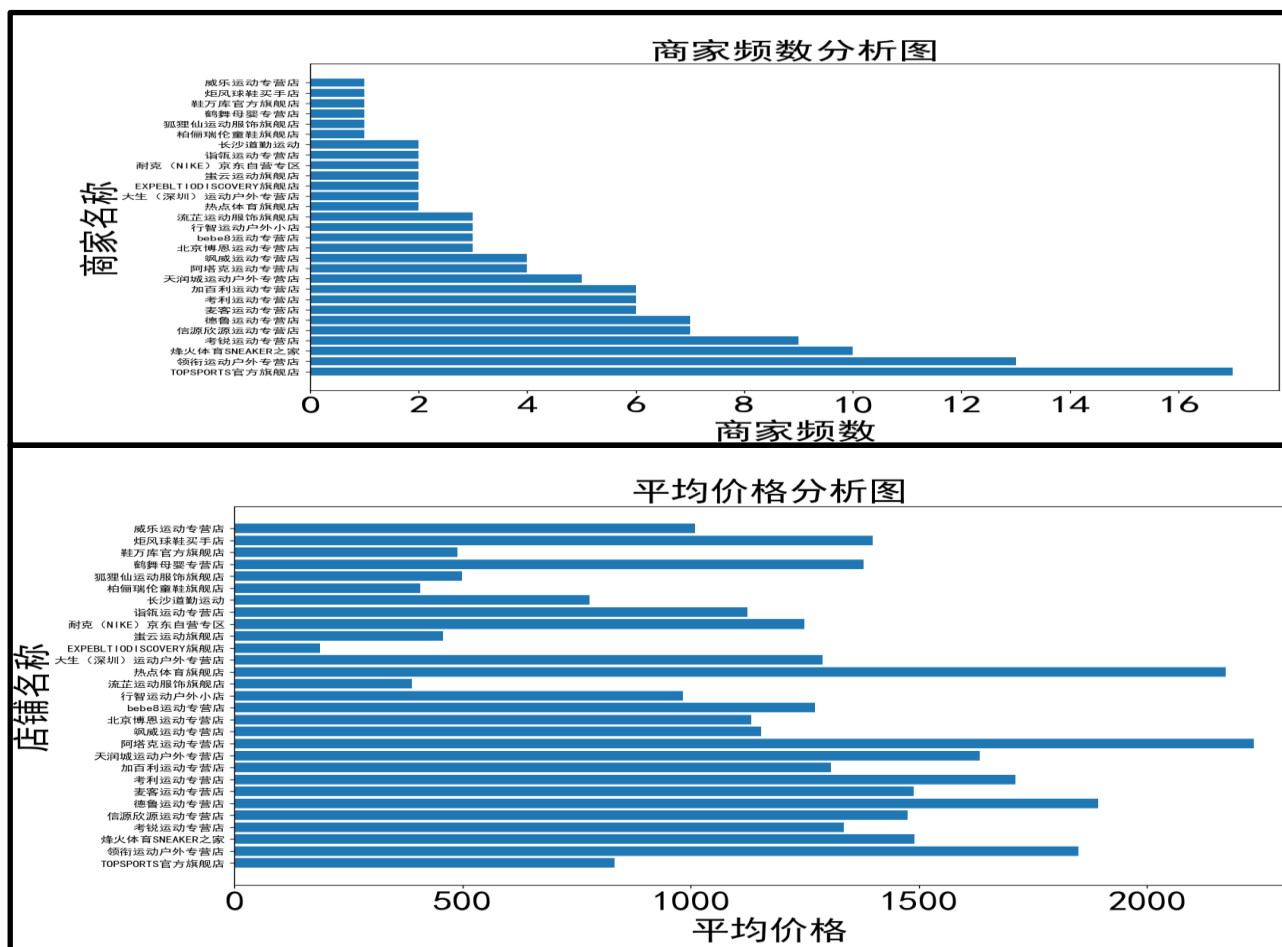


图 12 ‘AJ1’ 休闲鞋 商家平均价格与频数分析图

小结：

注：我们认为京东购物网站推荐的商品参数都是最受消费者欢迎的。

发现 TOPSPORTS 官方旗舰店官方推荐度最高（频数高）并且平均价格较低在 1000 元以下，因此推荐那位鞋码 42 持有 1000 元的朋友去 TOPSPORTS 官方旗舰店购买 AJ1 鞋。

四、关键问题及解决方法

1、数据爬取时遇到的问题与解决办法（王岩）

（1）打破网页反爬取障碍

刚开始按照最基础的网页访问方法如下：

```
import requests
if __name__ == '__main__':
    url = 'https://search.jd.com/Search?keyword=AJ1&enc=utf-8' # 目标网站地址
    response = requests.get(url) # 发送请求访问网站
    html = response.text # 获取网页源码
    print(html) # 将源码打印到控制台
```

这对于没有反爬取功能的网页是可以成功获取网页源码的，但是京东购物网站却会拦截我们的访问请求：

```
>>>
== RESTART: C:/Users/yeahamen/AppData/Local/Programs/Python/Python310/dawd.py ==
<script>window.location.href='https://passport.jd.com/uc/login'</script>
>>>
```

图 13 浏览器拦截代码运行-提示使用浏览访问作示意图

如何解决这个问题呢？首先思考：我们采用浏览器游客模式访问此网站时可以查看网页源码等数据，但利用 python 语言发动这个请求却不行，如果能修改 python 代码，使得其运行时模拟浏览器游客模式访问此网站就可以成功爬取数据了。只需在 request 函数中加一个‘附加条件’：‘headers=User-Agent...’。以 User-Agent 打头的代码块在搜索命令发起后的：开发者工具-网络-点击 search 条目-点击右侧标头-最后一行 User-Agent，完善后的代码如下：

```
import requests
if __name__ == '__main__':
    url = 'https://search.jd.com/Search?keyword=Leonard1&enc=utf-8' # 目标网站地址
    # 模拟浏览器访问
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/101.0.4951.64 Safari/537.36 Edg/101.0.1210.53'
    }
    response = requests.get(url, headers=headers) # 请求访问网站
    html = response.text # 获取网页源码
    print(html) # 将源码打印到控制台
```

然后直接获取到网页源码：

>

九

31

```
for k in range(0, 5): #一共爬取 5 次即 5 个页面
    i = 3+k*2 #页码每次加 2
    j = 56+k*60 #商品数量每次加 60
    keyword = 'AJ1' # 搜索关键字
    search_url='https://search.jd.com/Search?keyword=%s&suggest=1.his.0.0&wq=AJ1&pvid=65892364c2604d5897754ab21bed6d22&page=%d&s=%d&click=1'%(keyword,i,j)
```

得到的结果如下：

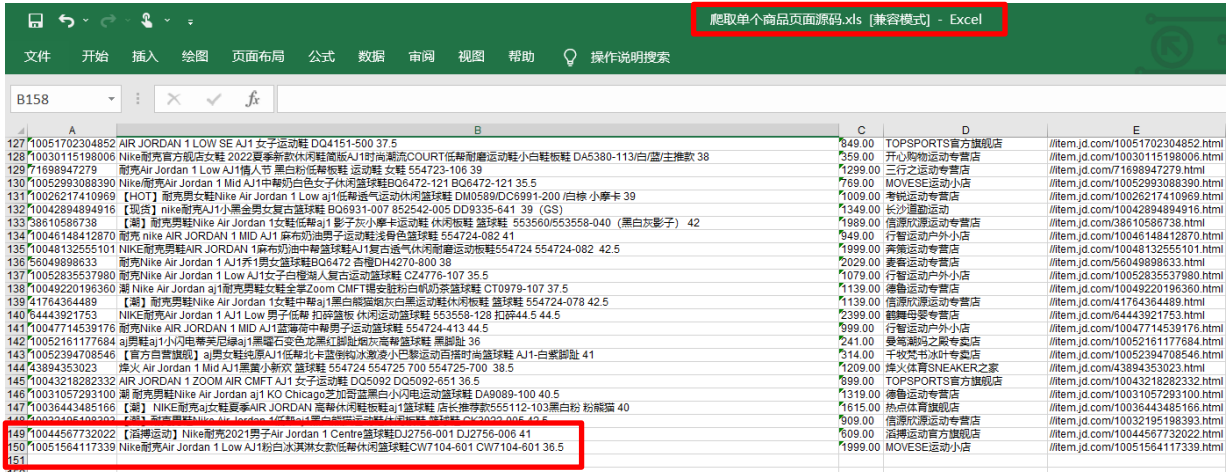


图 17 爬取五个页面商品信息 Excel 文件示意图

小结：

网页数据爬取过程中遇到的困难可以通过查阅资料、向水平更高的人请教学习，同时也要自己开动脑筋思考，善于发现（不同页面网页链接之间的区别），从而在已有的代码基础上稍加改动（添加循环格式）便能完善代码，实现爬取目的。

2、数据分析时遇到的问题与解决办法（郭文静）

（1）鞋码的正确提取

商品名称部分，最后的号码代表的是鞋码，在把鞋码单独提取到一个文档里得到一系列鞋码的过程中由于一些鞋码有些是单独的数字，有些是数字带有一个‘码’字或‘男’字，刚开始提取的都是最后的数字之后的，得到的后面有些带有‘码’或‘男’字，即图 1。

1	42.5	16	40.5
2	40	17	36
3	36.5	18	36
4	42	19	42
5	41	20	37.5
6	39	21	27
7	42	22	42
8	37.5码	23	39
9	40（男）	24	38
10	43	25	36
11	43	26	37.5
12	35	27	42
13	38	28	36
14	36	29	27
15	39	30	43

图 18 鞋码中带有其他字符-Excel 文件截图

通过百度搜索以及查找书籍发现了函数 findall 提取字符串中浮点数的功能，之后建立一个空列表，遍历每一个商品名称，取出结尾部分（数字和文字），并只提取数字部分，如：

```
elif re.findall('(\d\d.\d)码', good[i]) != []:
```

它可以匹配（数字+数字+除\n 以外任意字符+数字）+码，但只保留括号内的内容，具体代码如下所示

```
if re.findall('\d\d.\d$', good[i]) != []: #遍历每一个商品名称，下面 sss 指的是一些字符
    xiema.append(re.findall('\d\d.\d$', good[i]))
elif re.findall('\d\d$', good[i]) != []: #匹配以实数结尾的数字如：sssss42
    xiema.append(re.findall('\d\d$', good[i]))
elif re.findall('(\d\d.\d)码', good[i]) != []: #匹配（数字+数字+除\n 以外任意字符+
    xiema.append(re.findall('(\d\d.\d)码', good[i]))
elif re.findall('\s\d\d', good[i]) != []: #匹配空白字符/制表符/回车（这里是空白符）+数
    xiema.append(re.findall('\s(\d\d)', good[i]))#匹配时+空白符，添加到列表时只添加数
字
```

修改之后得到运行结果，得到新的列表，即图 2

		A	140	44
1	42.5	141	38	
2	40	142	37.5	
3	36.5	143	36	
4	42	144	36	
5	41	145	40.5	
6	39	146	37.5	
7	42	147	36	
8	37.5	148	36.5	
		149		

图 19 无其他字符的鞋码提取列表及 Excel 截图图

(2) 将从原文档中提取的‘商家’数据写入单独文档

在分析相同名称的店铺出现的次数，将它统计好：一列店铺名称一列出现次数，最后可以得到运行结果，即得到相同名称的店铺出现的次数，却不知道怎么写进文档中，即只得到了 Counter 固定数据类型：

```
Counter({'TOPSPORTS官方旗舰店': 17, '领衔运动户外专营店': 13, '烽火体育SNEAKER之家': 10, '考锐运动专营店': 9, '信源欣源运动专营店': 7, '德鲁运动专营店': 7, '加百利运动专营店': 6, '考利运动专营店': 6, '麦客运动专营店': 6, '天润城运动户外专营店': 5, '飒威运动专营店': 4, '阿塔克运动专营店': 4, '北京博恩运动专营店': 3, '行智运动户外专营店': 3, '流芷运动服饰旗舰店': 3, 'bebe8运动专营店': 3, '大生(深圳)运动户外专营店': 2, '诣领运动专营店': 2, '耐克(NIKE)京东自营专区': 2, '长沙道勤运动': 2, '蜚云运动旗舰店': 2, '热点体育旗舰店': 2, 'EXPEBLTIODISCOVERY旗舰店': 2, '爱萌仔旗舰店': 1, '阔茂运动户外旗舰店': 1, '君冈运动服饰旗舰店': 1, 'nan': 1, 'YYsports旗舰店': 1, '裳淇运动户外旗舰店': 1, 'Sneaker Club俱乐部': 1, '皓洋潮汇运动专营店': 1, '黑石旗舰店': 1, '任行运动户外小店': 1, '千牧朗沃道专卖店': 1, 'MOVESE运动小店': 1, '君悦运动专营店': 1, '旺热运动装备专营店': 1, '沙筱奕母婴旗舰店': 1, '炬风球鞋买手店': 1, '大格家居专营店': 1, '逗哒运动旗舰店': 1, '志胜天官方旗舰店': 1, '智凯运动专营店': 1, '威尔运动专营店': 1, '大锤运动户外小店': 1, '鞋万库官方旗舰店': 1, '狐狸仙运动服饰旗舰店': 1, '柏伽瑞伦童鞋旗舰店': 1, '瑾巡运动旗舰店': 1, '退仙旗舰店': 1})
```

通过上网搜索以及队员的提示，我发现运行结果是一个元组，元组由键值对组成，想要得到一列店铺名一列出现次数，那么需要编写程序：第一列只读键，第二列只读值就可以了，如下列代码：

key = list(d.keys()) 可以提取字典 d 中的所有键

value = list(d.values()) 可以提取 d 中所有的“值”

具体代码如下所示：

```
c = Counter(list1) # 在一个数组内，遍历所有元素，将元素出现的次数记下来
d = dict(c) # 转化成字典
print(d)

# 提取字典中的两列值 key 是键值，value 是 cont【key】对应的值
key = list(d.keys()) # 提取字典 d 中的所有键
value = list(d.values()) # 提取 d 中所有的“值”

# 利用 pandas 模块先建立 DataFrame 类型，然后将两个上面的 list 存进去
result_excel = pd.DataFrame()
result_excel["商品名称"] = key
result_excel["频数"] = value

# 写入 excel
result_excel.to_excel('商品频率.xlsx')
```

修改之后成功将数据写入文档中，得到表格，如图 4。

C	D		
商店名称	频数		
TOPSPORTS官方旗舰店	17	流芷运动服饰旗舰店	3
领街运动户外专营店	13	热点体育旗舰店	2
烽火体育SNEAKER之家	10	大生（深圳）运动户外专营店	2
考锐运动专营店	9	EXPEBLTIDISCOVERY旗舰店	2
信源欣源运动专营店	7	蚩云运动旗舰店	2
德鲁运动专营店	7	耐克（NIKE）京东自营专区	2
麦客运动专营店	6	诣瓴运动专营店	2
考利运动专营店	6	长沙道勤运动	2
加百利运动专营店	6	柏俪瑞伦童鞋旗舰店	1
天润城运动户外专营店	5	狐狸仙运动服饰旗舰店	1
阿塔克运动专营店	4	鹤舞母婴专营店	1
飒威运动专营店	4	鞋万库官方旗舰店	1
北京博恩运动专营店	3	炬风球鞋买手店	1
bebe8运动专营店	3		
行智运动户外小店	3		

图 20 相同店铺出现次数的 Excel 示意图

小结：

根据问题 1：我知道了函数 findall 提取字符串中浮点数的功能，并且学会了在字符串中有数字和文字的情况下，如何设计程序取出结尾部分（数字和文字），并只提取数

字部分，收获颇深。

根据问题 2：我知道了如何将元组转换成字典，将字典的键值对分别写成两列，写入表格中，同时，遇到问题时，要多去查找资料，运用网络或者询问同学和老师能够事半功倍，如今网络上的资源很多，无论是通过智慧树还是中国大学慕课 MOOC 都是很好的平台让自己学习让自己成长，遇到不会的仔细思考后与同学或者老师讨论交流会更好的处理问题，也能了解到与自己想法不同思路不同处理问题的方法

3、数据可视化时遇到的问题与解决办法（张成业）

（1）函数不能识别中文和特殊符号，导致最终显示时中文以及部分符号会变成乱码有运行问题的代码如下：

```
#鞋码分布饼状图
#引入模板
import pandas as pd
import matplotlib.pyplot as plt
#获取路径
text=r"C:\Users\86178\Desktop\鞋码的分布情况 excel 文件.xls"
#读取文件
pe=pd.read_excel(text)
#获取数据
xx=pe['范围']
yy=pe['数量']
```

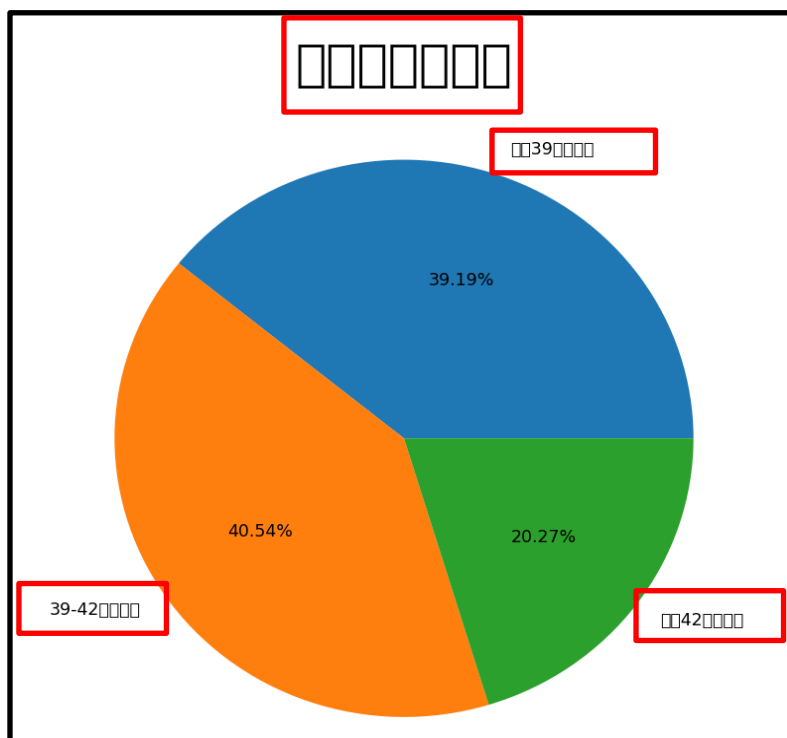


图 21 问题代码显示的鞋码分布

具体解决方法：经过多渠道询问调查，了解到需要相关正确的专业信息。在网上查阅相关资料后，发现要解决这个问题需要从软件的基础操作入手，需要使用 plt.rcParams 解决绘制过程中出现的问题。修改问题后的正确代码如下：

```
#鞋码分布饼状图
#引入模板
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import rcParams
plt.rcParams['font.sans-serif']=['Microsoft YaHei'] #显示中文标签,处理中文乱码问题
plt.rcParams['axes.unicode_minus']=False #坐标轴负号的处理
plt.axes(aspect='equal') #将横、纵坐标轴标准化处理，确保饼图是一个正圆，否则为椭圆
rcParams['font.family'] = 'simhei'
```

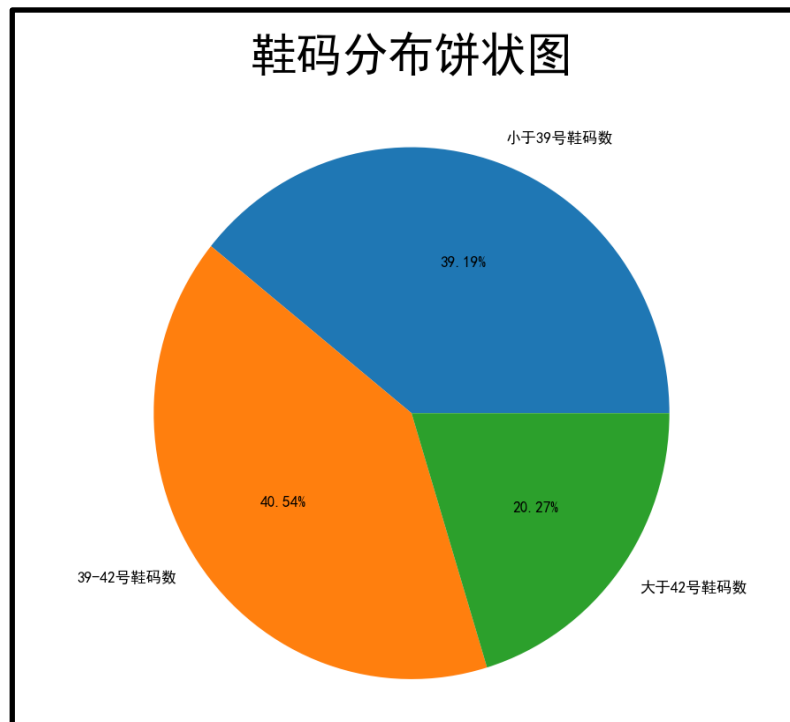


图 22 无乱码显示的鞋码分布饼状图

(2) 横纵标轴名称过小，显示时不太明显。

```
k=pe['词向量']
m=pe['词频']
#绘制图形
plt.xlabel('商家频数')#横轴名称
plt.ylabel('商家名称')#纵轴名称
plt.title('商家频数分析图')
plt.barh(k,m)
plt.show()
```

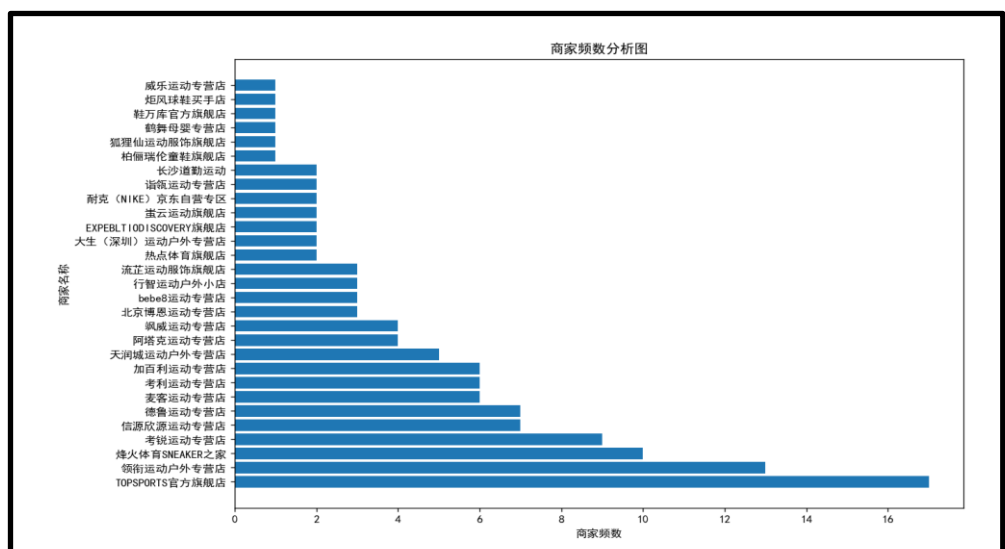


图 23 标题与横纵坐标名称显示过小的商家频数分析图

解决方法：在经过思考及查阅资料后，使用 `plt.x/ylabel(' ', fontsize)` 将出现的此

```
#绘制图形
plt.xlabel('商家频数',fontsize=30)#横轴名称
plt.ylabel('商家名称',fontsize=30)#纵轴名称
plt.title('商家频数分析图',fontsize=30)
plt.xticks(fontsize=30)
plt.barh(k,m)
plt.show()
```

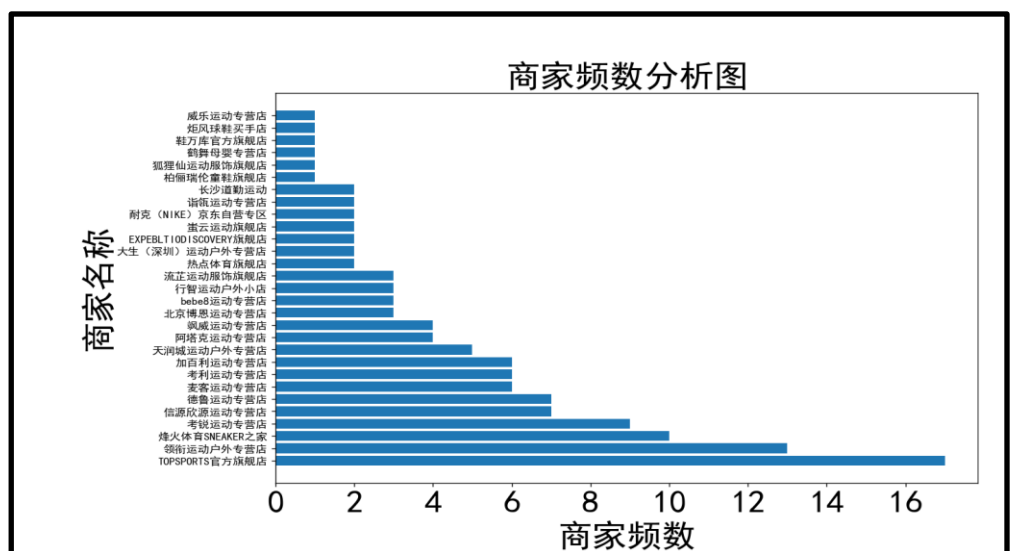


图 23 标题与横纵坐标名称显示恰当的商家频数分析图

五、数据处理与分析结果分析

综合上面所制作的饼状图、雷达图、条方图，可以得出以下结论：

（1）网站的鞋码推荐大小范围分布与中国社会成年人鞋码范围分布有很好地相关性。所以‘AJ1’休闲鞋适合所有成年人购买穿戴。男性朋友的 42 号穿鞋标准可以购买 AJ1 鞋。

(2) ‘AJ1’ 休闲鞋平均售价在 1000 元及以下的有：TOPSPORTS 官方旗舰店、行智运动户外小店、长沙道勤运动等店铺。可以推荐持 1000 元购鞋的朋友去这几个店铺挑选。

(3) 爬取到的所有店铺的平均价格分布在 1000-1500 的最多，因此朋友的 1000 元可能不够，建议他多准备 100-200 元。

(4) 综合网站所推荐各个店铺的频数排序与店铺对应的平均价格分析，发现 TOPSPORTS 官方旗舰店官方推荐度最高（频数最高）并且平均价格较低在 1000 元以下，因此推荐那位鞋码 42 持有 1000 元的男性朋友去 TOPSPORTS 官方旗舰店购买 AJ1 鞋。

六、学习总结与反思

王岩：

1、爬取京东购物网站的商品参数信息前，首先要把基本的网页设计内容搞懂。对于目标网页的网页链接构成[URL]、网页元素构成[ELEMENT]等都需要进行深入完整的理解，大部分的网站都有管理员所设置的‘反爬虫’功能，能否掌握利用用户代理[USER-AGENT]来模拟浏览器访问网站爬取信息是能否成功的关键。

2、成功模拟浏览器访问网站并获取网页源码之后[REQUEST]，需要在众多的网页元素中查询到与商品对应的代码条目[目的源码 Goodlist]。把所有 Goodlist 代码赋给一个列表，遍历此列表筛选出‘商家’、‘价格’、‘商品名称’等所需信息，把这些提取后的信息写入到 Excel 文件以供分析与可视化。

注：爬取网页时，针对京东购物网站，每一个搜索结果页面都会自动展示 30 件商品[代码可以在一个网页链接源码中直接筛选出来的信息]，用鼠标滚轮滑动至第三十件时，网站再次自动推送 30 件[代码会自动跳过这 30 件信息]。因此如果要爬取 150 件商品就要爬取 5 个页面，爬取多个页面就在于网页链接的循环使用，在网页链接最后有：

‘&page=3&s=56’ 只需在代码里更改‘page’与‘s’的数值即可轻松达到爬取多个网页的目的。

郭文静：

1、在编写程序的过程中，不一定自己遇到的问题就是很困难的解决不了的，也许问题就出在了很简单的地方，在生活中以及学习中要更加细心，要做好眼前的，也许很多问题出在自己会的但是粗心的地方，而不一定都是自己不会的地方，写完程序后多检查

两遍。

2、在学习的过程中，要及时的对自己不确定的或者不会的问题进行总结分析并做好备忘，以免真正需要的时候只记得曾经学过却忘记了具体操作，备忘要随时记，不可以偷懒，否则下次遇到问题时，就只记得自己曾经学过，但是具体内容全都忘了。

3、在编写程序过程中发现自己对第三方库的函数不了解，通过搜索查找后得知这些函数的用途。这也让我明白了在平时就应当锻炼自己能够快速查阅找到 python 各类第三方库的函数使用说明，以便自己以后写程序效率提高。

注：[1]耐心及细心是解决问题的关键。[2]要善于发现问题，打破平常的思路。[3]遇到自己不确定问题要记得及时做好备忘。

张成业：

1、通过本次作业发现自己对许多的 Python 基础知识掌握不够熟练，在实际操作时经常翻阅课本和上网查找相应的知识点才能解决问题，不过这也让自己更加熟练地运用 Python 解决实际问题。

2、在数据可视化过程中，需要制作饼图，柱状图，雷达图，绘制这些图形都一些共同点。在读写 Excel 文件时，选对相应的模块对 Excel 文件实现读写操作，本次作业中 Excel 文件的扩展名为“xlsx”所以选取 openpyxl 模块。绘制图形是需要引入 matplotlib 库。在绘制图形中一般会涉及汉字和特殊符号，这时就需要 plt.rcParams，图形中就会出现乱码的情况，有时需要改变字体的大小时 plt.xlabel(, fontsize)进行调节。

附：根据汇报 ppt 时老师提出的两个建议，进行了分析与优化。

(1) DataFrame 简化鞋码分布过程

利用 pandas 库中的 cut（分布分析）、pivot_table（交叉分析）函数大大简化了汇总鞋码分布的过程，核心代码仅需两行，如下：

```
import pandas as pd
df=pd.read_excel(r"鞋码写入到 excel 里面.xls")
bins=[0,38.5,42,50]
label=["小号<39", "39<中号<42", "大号>42"]
df['鞋码分布']=pd.cut(df.鞋码,bins,right=True,labels=label)
df_pt = pd.pivot_table(df[['鞋码', '鞋码分布']],index="鞋码分布",aggfunc='count')
```

添加上写入 excel 文档的代码段：

```
df_pt = dict(df_pt)
key = list(df_pt.keys()) # 提取字典 d 中的所有键
value = list(df_pt.values()) # 提取 d 中所有的"值"
result_excel = pd.DataFrame() # 利用 pandas 模块先建立 DataFrame 类型
result_excel["鞋码分布"] = key # 把 key、value 列表存进去
result_excel["分布情况"] = value
result_excel.to_excel('利用 cutpivot 函数整理鞋码分布.xlsx') # 写入 excel
```

得到分布结果：

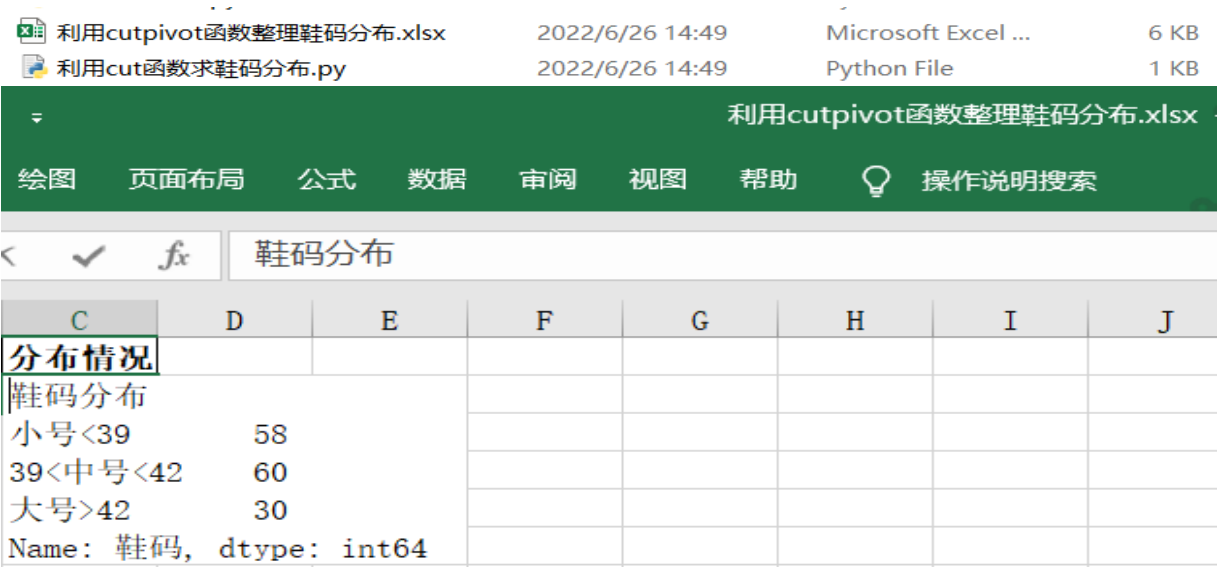


图 23 利用 cut、pivot 函数整理鞋码分布 Excel 文件截图

(2) 分析与商家（店铺）的平均价格与推荐频数相关性。

本学期刚好学习完课程：《概率论与数理统计》，分析两组数据的相关性经常用到皮尔逊相关值即相关系数，公式如下：

$$\rho_{xy} = \frac{cov(x,y)}{\sqrt{DX}\sqrt{DY}} = \frac{EXY - EXEY}{\sqrt{DX}\sqrt{DY}}$$

我把两组数据放在下面：

商家	TOPSPORT	领衔运动	烽火体育	考锐运动	信源欣源	德鲁运动	麦客运动	考利运动	加百利运	天润城运	阿塔克运
频数	17	13	10	9	7	7	6	6	6	5	4
平均价格	833.706	1849.77	1491	1335.67	1474.71	1893.29	1489	1712.33	1307.33	1633.2	2234
商家	飒威运动	北京博恩	bebe8运动	行智运动	流芷运动	热点体育	大生（深	EXPEBLTI	蚩云运动	耐克（NI	诣瓴运动
频数	4	3	3	3	3	2	2	2	2	2	2
平均价格	1154	1132.333	1272.333	982.3333	389.1667	2172	1289	188	457	1249	1124
商家	长沙道勤	柏佰瑞伦	狐狸仙运	鹤舞母婴	鞋万库官	炬风球鞋	威乐运动	智凯运动	三行之运	瑾巡运动	志胜天官
频数	2	1	1	1	1	1	1	1	1	1	1
平均价格	778.94	408	499	1379	489	1399	1009	3011	1199	439	1039
商家	逗哒运动	大格家居	大锤运动	裳淇运动	Sneaker（	沙筱奕母	旺热运动	君悦运动	爱萌仔旗	MOVESE运	阔茂运动
频数	1	1	1	1	1	1	1	1	1	1	1
平均价格	855.2	549	679	668	1069	297	1599.99	1559	208	1179	258
商家	千牧朗沃	君闵运动	服饰旗舰店	YYsports	任行运动	黑石旗舰	皓洋潮汇				
频数	1	1	1	1	1	1	1				
平均价格	376	468	1129	899	799	1179	388				

图 24 商家频数、平均价格数据展示图

```

import pandas as pd
import openpyxl
import math
i=0
sum_pinshu=0
sum_jiage=0
sum_pinshu_jiage=0
EX_pinshu=0
EX_jiage=0
DX_pinshu=0
DX_jiage=0

wb=openpyxl.load_workbook(r"商品频率.xlsx") #读取文件商品频率
ws=wb.active #获取工作表索引值，无更改则一直工作这一张
data1 = pd.read_excel(r"商品频率.xlsx") #读取文件‘商品频率’给 data1
pinshu = data1["频数"] #读取 data1 ‘商家’ 列（49）
jiage = data1["平均价格"]
pinshu = list(pinshu)
jiage = list(jiage)
for i in range(0,50):
    sum_pinshu+=pinshu[i]
    sum_jiage+=jiage[i]
    sum_pinshu_jiage+=jiage[i]*pinshu[i]
EX_pinshu = sum_pinshu/50#频数期望
EX_jiage = sum_jiage/50#价格期望
EXY_pinshu_jiage = sum_pinshu_jiage/50
for i in range(0,50):
    sumDX_pinshu = (pinshu[i]-EX_pinshu)**2

```



```

        sumDX_jiage = (jiage[i]-EX_jiage)**2
    DX_pinshu = math.pow(sumDX_pinshu/50,1/2)#频数方差
    DX_jiage = math.pow(sumDX_jiage/50,1/2)#价格方差
    pxy=(
    EXY_pinshu_jiage-
EX_pinshu*EX_jiage)/(math.pow(DX_pinshu,1/2)*math.pow(DX_jiage,1/2))#相关系数
    print('频数期望=',EX_pinshu,'频数方差=',DX_pinshu)
    print('价格期望=',EX_jiage,'价格方差=',DX_jiage)
    print('EXY=',EXY_pinshu_jiage)
    print('相关系数 pxy=',pxy)

```

得到的计算结果如下：

```

===== RESTART: C:/Users/yeahamen/Desktop/新建文件夹 (6)/求平均价格与频数的期望和方差.py =
=====
频数期望= 2.94 频数方差= 0.2743574311003804
价格期望= 1081.6861022624432 价格方差= 13.762263398784123
EXY= 3759.6714
相关系数pxy= 298.2365246327246

```

图 24 频数、价格的期望、方差、相关系数计算结果图

发现 $|\rho_{xy}| \gg 1$, 因此商家（店铺）平均售价与出现的频数相关性不强，男性朋友按照前面的结论购买‘AJ1’休闲鞋即可。

最后感谢郭老师在本次课程教学中给予的宝贵的指导，祝老师身体健康，工作顺利！