華中科技大学

《大数据导论》课程报告

题目： 京东服装数据—测试分析组

课程名称： 大数据导论

专业班级：

同小组成员：

同大组成员：

指导教师：

报告日期：

计算机科学与技术学院

目 录

1 任务简介

1.1 任务分工

任务：

(1) 提取数据组爬取的数据；

(2) 整理数据；

(3) 数据关联分析；

(4) 数据可视化；

(5) 结果分析。

1.2 详细分工

(1)

对数据集进行统计处理：通过代码形式对数据进行统计，使用 matplotlib 绘图库对统计后数据进行整理与可视化。

提取商品名中可用信息：通过代码，提取原数据中未细化的数据，并对其进行统计分析。

组装各数据属性的关联性：通过代码实现数据关联，将多种数据进行关联分析，最终进行可视化绘图。

(2)

对数据进行细化分析：总览数据集，讲包含信息量较多的属性再进行信息提取，以便后续数据关联性分析；

数据关联性分析：将处理后的数据属性间，根据制作的属性图，进行关联分析，绘制相关关系图，方便后续阅览与查看；

整理最终结果：参考可视化数据，对数据进行详细论述分析，并针对应用层面，对数据进行进一步计算处理，提供合理化建议，使数据具有参考价值。

2 测试分析过程

2.1 数据细化

总览数据集：

商品名	价格	店铺名	销售量	广告
威廉爵驰休闲裤男20年夏季新款运动卫裤子男士	88.00	威廉爵驰旗舰店	4800+	关注店铺即可享受粉丝优惠价
威廉爵驰休闲裤子男生阔腿直筒商务休闲小西裤	79.00	威廉爵驰旗舰店	4000+	关注店铺即可享受粉丝优惠价
威廉爵驰运动裤男休闲卫裤子男20年夏季新款男	88.00	威廉爵驰旗舰店	1.7万+	[面料升级，不起球，起球免费
威廉爵驰短袖t恤男纯色潮牌潮流ins五分原宿港	68.00	威廉爵驰旗舰店	7000+	关注店铺即可享受粉丝优惠价
威廉爵驰短袖t恤男潮牌潮流ins五分原宿港风百	58.00	威廉爵驰旗舰店	3100+	关注店铺即可享受粉丝优惠价
威廉爵驰休闲裤男20年夏季新款运动卫裤子男士	88.00	威廉爵驰旗舰店	500+	关注店铺即可享受粉丝优惠价

图 2-1 原数据集

我们发现，商品名一栏里，包含更多的可用信息，如款式、码数、性别、季节性等要素，于是我们对商品名进行进一步细化工作。

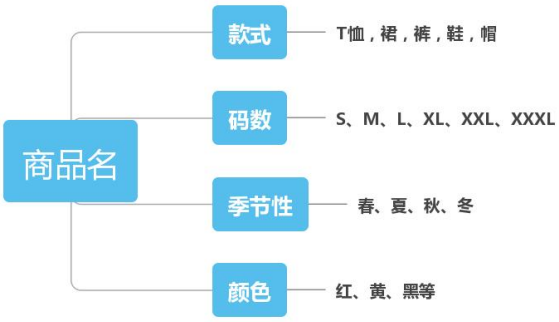

图 2-2 细化商品名属性

2.2 数据提取

通过细化，我们得到了如下属性：款式、码数、颜色、季节性、价格、店铺名、销售量以及广告。于是我们对数据进行了相应的提取：

①服装款式：鞋、裙、恤、裤、帽、包。

②服装码数：S、M、L、XL、XXL、XXXL、XXXXL、XXXXXL；

鞋码：34-46。

③季节性：春、夏、秋、冬。

④颜色：白、黑、灰、红、绿、黄、紫、蓝。

⑤款式：长、短；大、小；休闲、运动；男、女。

2.3 数据关联分析

属性之多对后续的分析工作产生巨大阻碍，对数据的关联分析极大降低了非必要数据对分析的干扰。由于上述属性有 2^7 种关联方法，故我们只挑选其中具有实际价值的数据进行分析。

①分析“鞋、裙、恤、裤、帽”的销量；

②分析服装码数的销量，鞋码数的销量；

③不同季节的商品数，分析夏季及其他季节的比例；

④分析各商品价格的平均值；

⑤选取 5 个服装店进行价格比较；

⑥男包、女包的平均价格，销售量比例以及商品数比例。

对于以上分析，我们选取了两组分析案例：包-性别-价格-销量；服装-款式-季节性-商品数。此仅为基础关联，我们还将对其进行进一步细分，于是有了以下 4 个分析实例：（1）各季服装的比较；（2）服装码式；（3）服装类别；（4）包类分析。

2.4 代码实现

接下来，进行代码操作。我们将数据统计分成了两个模块，第一，实现数据统计；第二，实现 matplotlib 绘图操作。相关步骤代码截图如下，详细见附录。

```
5.  import matplotlib.pyplot as plt
6.
7.  xl = xlrd.open_workbook(r'C:\Users\LH2019\Desktop\jd_data.xlsx') # 读取数据表
    格文件的地址
8.  mysheet = xl.sheets()[0]
9.
10. spr_aut = 0
11. summer = 0
12. winter = 0
13. row = 1
14. while row <= 495376:
15.     data = mysheet.cell(row,0).value
16.     if '夏' in data:
17.         summer += 1
18.     if '春' in data or '秋' in data:
19.         spr_aut += 1
20.     if '冬' in data:
21.         winter += 1
22.     row += 1
23.
24. def autolabel(rects):
25.     for rect in rects:
26.         height = rect.get_height()
27.         plt.text(rect.get_x()+rect.get_width()/2.- 0.2, 1.03*height, '%s' %
    int(height))
```

图 2-3 代码-数据统计

```
28.
29. print('夏季服装所占比例：', summer*100/(summer+spr_aut+winter),'%')
30. print('反季服装所占比例：', winter*100/(summer+spr_aut+winter),'%')
31. name_list = ['Summer', 'Spring/Autumn', 'Winter']
32. num_list = [summer, spr_aut, winter]
33. autolabel(plt.bar(range(len(num_list)), num_list, color='rgb', tick_label=na
    me_list))
34. plt.show()
35.
```

图 2-4 代码-matplotlib 绘图

```
PS C:\Users\LH2019> & D:/SOFTWARE/python/anaconda3_2020/envs/tensorflow/python.exe c:/Users/LH2019/Desktop/
夏季服装所占比例： 68.69604063519644 %
反季服装所占比例： 4.142499617440753 %
```

图 2-5 代码-运行结果

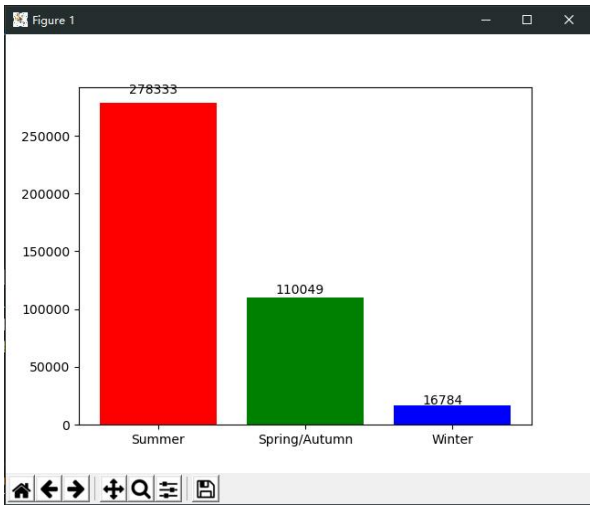

图 2-6 代码-绘图结果

3 成果展示

（1）各季服装的比较

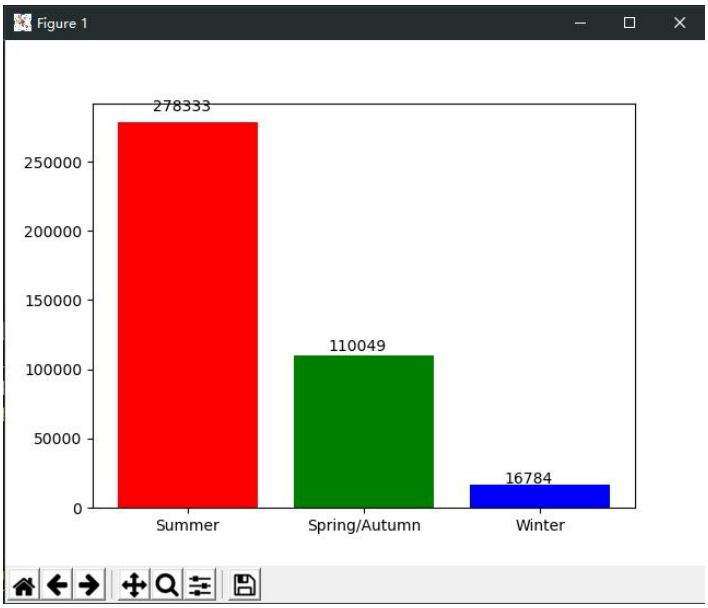

图 2-7 各季服装比较

分析：夏季：冬季=17：1；夏季：非夏季=2：1。夏季的服装商品销售平台依然有 1/3 的非夏季服装商品，且价格低廉。可以看出，京东服装平台的服装根据当前季节做出明显调整，符合季节性消费的原则。但非夏季服装的比例也占据一定比例，故销售平台为照顾特殊顾客提供了更多的选择。

建议：对于销售服装的店家，可以考虑为非应季服装腾出一部分席位，一方面处理余货，一方面满足了部分特殊顾客的需求，增加利润收入。建议商品数应季：非应季比例为 10:1。

（2）服装码式

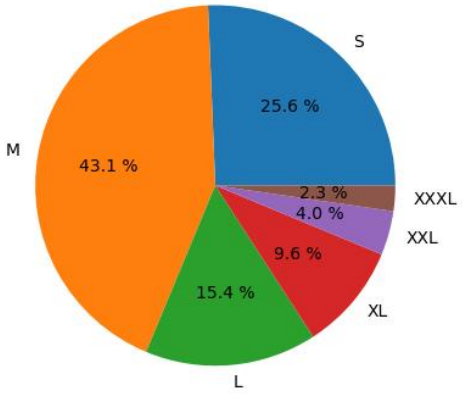

图 2-8 服装码式

分析：M 码超四成，比其他码数占比都要大。商品码数占比从大到小排列为：M，S，L，XL，XXL，XXXL。且 XXL 与 XXXL 比例极小。

建议：特殊服装除外，一般而言，服装商家进购服装商品时，建议将购置比例为最大定为 M 码。并降低 XXL 与 XXXL 码的比例，经计算，建议比例为 15:1。

（3）服装类别

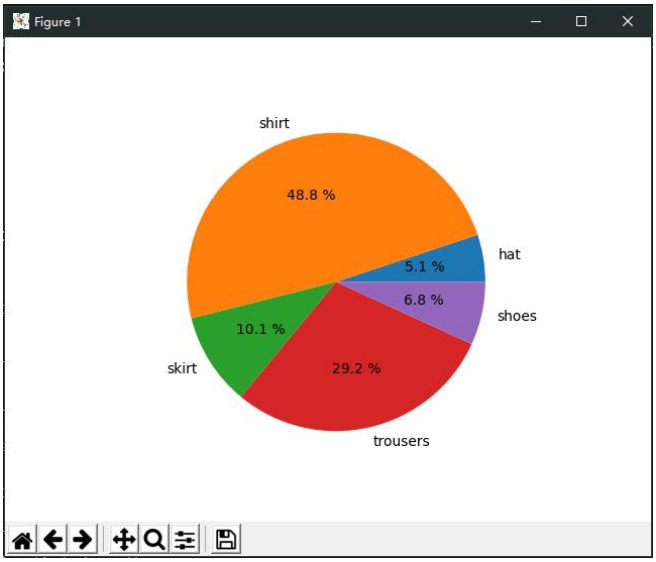

图 2-9 服装类别

分析：做该比例分析，是为解决多样化服装商铺的服装类别购置比例问题。由右图可以得出以下结论：夏季 T 恤的商品比例最大，帽与鞋的比例较小。比例由大到小排列为：恤，裤，裙，鞋，帽。

建议：此建议仅争对多样化店铺。在夏季中，建议将 shirt 的购置比例升高，考虑到帽与鞋为多样化商铺的补充，故建议仅做微调，按自身情况进行自主调试。建议比例：shirt：其他 = 1：1。

（4）包类分析

```
PS C:\Users\LH2019> & D:/SOFTWARE/python/anaconda3_2020/env
男包比例：  47.74023694602896 %
女包比例：  52.25976305397104 %
男包平均价格：  266.04064031862754
女包平均价格：  222.086742233418
```

图 2-10 包类分析 1

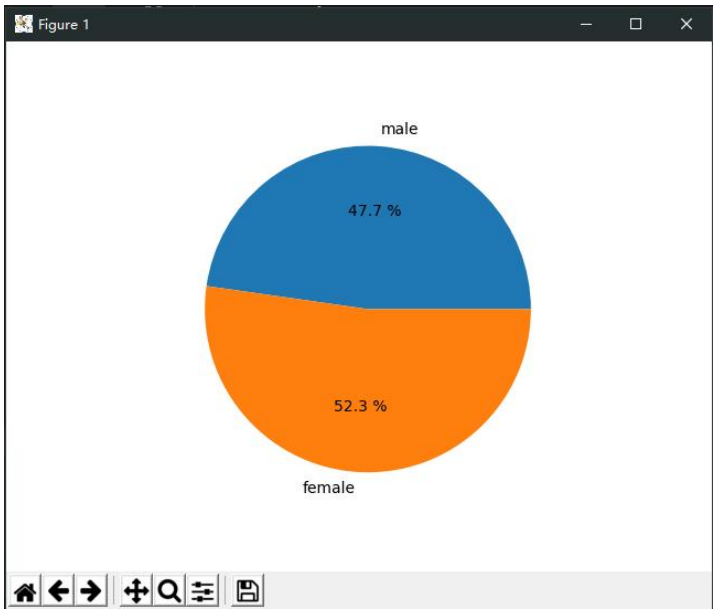

图 2-11 包类分析 2

分析：惊奇的发现，男包与女包的商品数量比接近 1：1。但男包与女包的平均价格却是男包价格大于女包。可以推测，网络平台女包种类数量的多样性使得各价格区间的女包都有一定的占比。而因男士相对不注重包的外观，男包种类相对较少，故包的价格单一，且集中与某一区间。

建议：

1.对于求稳型店家，建议按包类趋势购置商品，建议比例 1：1，价格置为普通区间价格。

2.对于突破型店家，建议引进更加多样化的男包，填补男包款式空缺，可能会有巨大经济效益。经计算，对于男包，建议男包新款式与普通款式男包比例为：24:1。

4 工具说明

编程环境：Python 3.7.7
绘图环境：matplotlib 3.2.1

5 心得体会

(1)

在这次大数据的项目中，我负责的测试分析级。在这次大数据的项目中，我负责的数据分析任务从 4 个角度对京东的服装数据进行了统计和分析，对当前服装市场的现状有了更加深入的了解，真切体会到了大数据中所蕴含的巨大价值。通过程序的实践，我也加深了对 Python 和其 matplotlib 绘图库的理解和掌握，增强了通过科学技术手段解决实际问题的能力。

(2)

在这次大数据的项目中，我负责的测试分析级。通过对数据的整理和分析，不仅锻炼了我的耐力，还锻炼了我的数据分析能力与数据判断能力。在冗杂的数据中，如何抓住关键数据至关重要。要有敏锐的联想力，如何联立数据是关键。

另外，此次项目是小组合作完成，锻炼了每个人的合作能力，懂得了如何分工，如何处理过程问题等。在测试分析级中，我们将任务细化，并进行合理分工，这对今后的学习与工作有十分重要的意义。

最后，我也感受到了大数据的魅力。在众多数据中获得可视化结果，并分析出价值建议，收获成就感的同时也造福我们的社会，这是大数据带给我最大的惊喜。

6 附录

```
# 本程序功能：1、利用 matplotlib 绘图库画出当前京东各季服装数量
#             2、输出夏季及反季服装所占的比例

import xlrd
import matplotlib.pyplot as plt

xl = xlrd.open_workbook(r'C:\Users\LH2019\Desktop\jd_data.xlsx') # 读取数据表格文件的地址
mysheet = xl.sheets()[0]

spr_aut = 0
summer = 0
winter = 0
row = 1
while row <= 495376:
    data = mysheet.cell(row,0).value
    if '夏' in data:
        summer += 1
    if '春' in data or '秋' in data:
        spr_aut += 1
    if '冬' in data:
        winter += 1
    row += 1

def autolabel(rects):
    for rect in rects:
        height = rect.get_height()
        plt.text(rect.get_x()+rect.get_width()/2.- 0.2, 1.03*height, '%s' % int(height))

print('夏季服装所占比例：', summer*100/(summer+spr_aut+winter),'%')
print('反季服装所占比例：', winter*100/(summer+spr_aut+winter),'%')
name_list = ['Summer', 'Spring/Autumn', 'Winter']
num_list = [summer, spr_aut, winter]
autolabel(plt.bar(range(len(num_list)), num_list, color='rgb', tick_label=name_list))
plt.show()

```

```
# 程序功能：统计出京东在售商品中各码数服装所占比例，以饼状图展示

import xlrd
import matplotlib.pyplot as plt

xl = xlrd.open_workbook(r'C:\Users\LH2019\Desktop\jd_data.xlsx') # 读取数据表格文件的地址
mysheet = xl.sheets()[0]

s = 0
m = 0
l = 0
xl = 0
xxl = 0
xxxl = 0
row = 1
while row <= 495376:
```

```
    data = mysheet.cell(row,0).value
    if 'S' in data:
        s += 1
    if 'M' in data:
        m += 1
    if 'XXXL' in data or '3XL' in data:
        xxxl += 1
    elif 'XXL' in data or '2XL' in data:
        xxl += 1
    elif 'XL' in data:
        xl += 1
    elif 'L' in data:
        l += 1
    else:
        pass
    row += 1

# 画饼状图
name_list = ['S', 'M', 'L', 'XL', 'XXL', 'XXXL']
num_list = [s, m, l, xl, xxl, xxxl]
# 保证圆形
plt.axes(aspect=1)
plt.pie(x=num_list, labels=name_list, autopct='%3.1f %%')
plt.show()
```

```
# 本程序功能：算出帽、恤、裙、裤、鞋在京东在售服装中所占的比例，最终用饼状图呈现

import xlrd
import matplotlib.pyplot as plt

xl = xlrd.open_workbook(r'C:\Users\LH2019\Desktop\jd_data.xlsx') # 读取数据表格文件的地址
mysheet = xl.sheets()[0]
hat = 0
shirt = 0
skirt = 0
trousers = 0
shoes = 0
row = 1
while row <= 495376:
    data = mysheet.cell(row,0).value
    if '帽' in data:
        hat += 1
    if '恤' in data:
        shirt += 1
    if '裙' in data:
        skirt += 1
    if '裤' in data:
        trousers += 1
    if '鞋' in data:
        shoes += 1
    row += 1

# 画饼状图
name_list = ['hat', 'shirt', 'skirt', 'trousers', 'shoes']
num_list = [hat, shirt, skirt, trousers, shoes]
# 保证圆形
plt.axes(aspect=1)
plt.pie(x=num_list, labels=name_list, autopct='%3.1f %%')
```

```
plt.show()
```

```
# 程序功能：1、得出京东商品中男包女包所占的比例，并画出饼图以直观表示
#           2、统计京东商品中男包女包的平均价格

import xlrd
import matplotlib.pyplot as plt

xl = xlrd.open_workbook(r'C:\Users\LH2019\Desktop\jd_data.xlsx') # 读取数据表格文件的地址
mysheet = xl.sheets()[0]

male = 0
price_male = 0
female = 0
price_female = 0
row = 1
interference_list = ['包臀','面包','全包','包头','包边','包邮','表情包','包装','礼包','包芯','包裙']
while row <= 495376:
    data = mysheet.cell(row,0).value
    if data != '' and mysheet.cell(row,1).value != '0.05/千字' and mysheet.cell(row,1).value != '暂无报价' and mysheet.cell(row,1).value != '价格':
        price = float(mysheet.cell(row,1).value)
        if '包' in data:
            flag = 1
            for item in interference_list:
                if item in data:
                    flag = 0
            if flag == 1:
                if '男' in data:
                    male += 1
                    price_male += price
                if '女' in data:
                    female += 1
                    price_female += price
    row += 1

print('男包比例：', male*100/(male+female),'%')
print('女包比例：', female*100/(male+female),'%')
print('男包平均价格：', price_male/male)
print('女包平均价格：', price_female/female)

# 画饼状图
name_list = ['male', 'female']
num_list = [male, female]
# 保证圆形
plt.axes(aspect=1)
plt.pie(x=num_list, labels=name_list, autopct='%3.1f %%')
plt.show()
```