任务 1 数据导入与预处理

任务 1.1 探查数据质量并进行缺失值和异常值处理

1.1.1 数据结构总览

查看数据集项数，发现数据集 data1.csv,有 4341 项，5 列；数据集 data2.csv,有 519367 项，14 列；数据集 data3.csv,有 43156 项，6 列

1.1.2 检查重复值

通过去重操作发现三个数据集均无重复项

1.1.3 数据内容总览

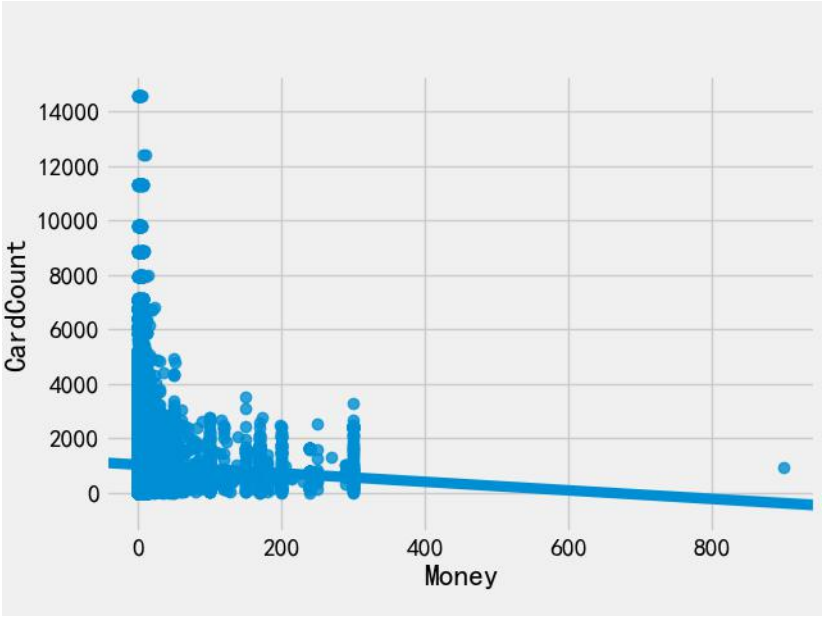
查看数据信息 info(),发现 data1 和 data3 中均无缺失值 ，data2 中 termSerNo、conOperNo 存在较大量的缺失值,因为这两项数据对后续分析无影响故直接过滤

1.1.4 数据分布总览

通过对数据 Describe，查看数据的均值，最大值，最小值以及方差等数据特征，观察到 data1 和 data3 中的特征值均较为合理，data2 中的 Money、FundMoney、Surplus 以及 CardCount，均存在和样本群体偏离程度较大的数据，会影响后序模型的性能

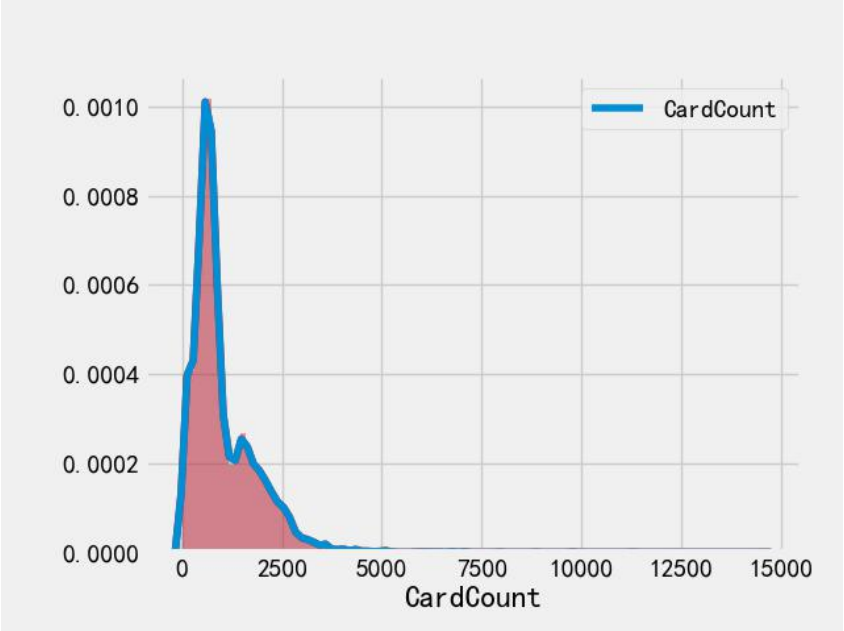
1.1.5 消费金额和消费次数

观察消费金额和消费次数的散点图，发现数据中具有一定数量的离群点，将其过滤



1.1.6 观察 CardCount 特征的分布情况

通过 distplot 和 kdeplot 绘制柱状图观察 CardCount 特征的分布情况，属于长尾类型的分布，这说明了有很多消费次数过多且超出正常范围。



任务 2 食堂就餐行为分析

任务 2.1 绘制各食堂就餐人次的占比饼图，分析学生早中晚餐的就餐地点，是否有显著差别

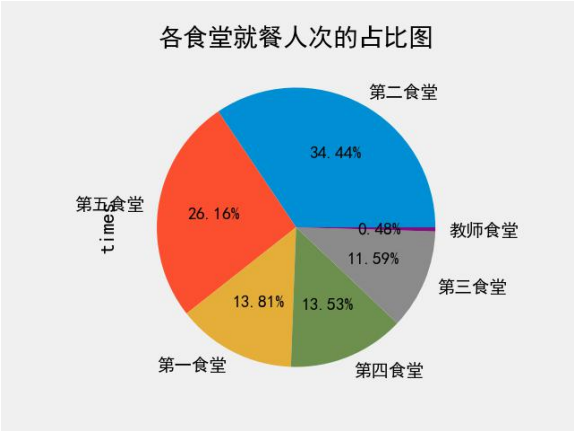


图 1

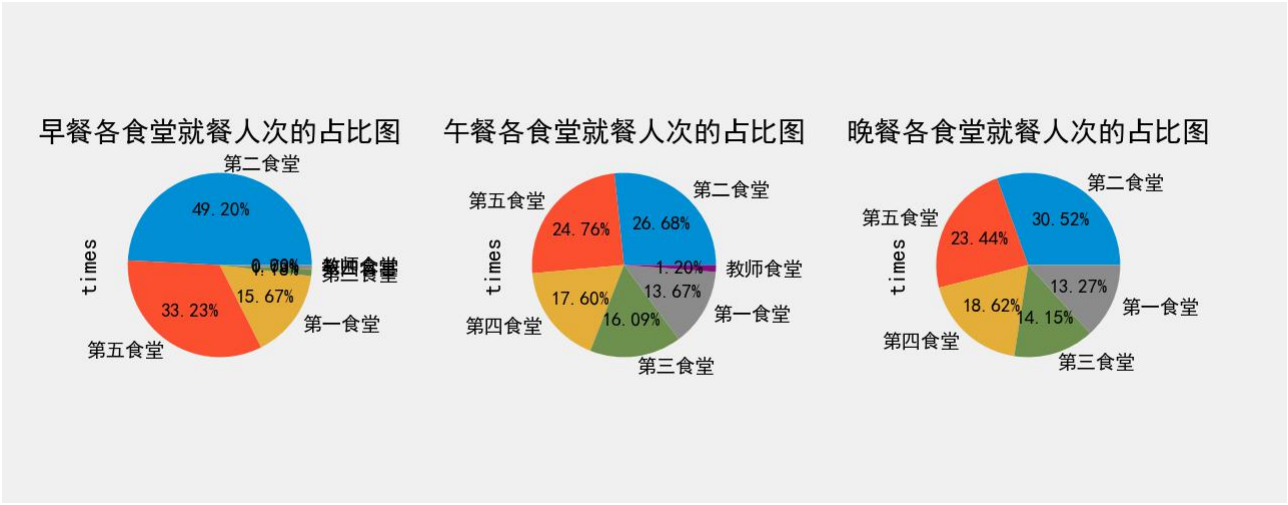


图 2，图 3，图 4

根据图 1，综合早、午、晚三餐学生的就餐地点来看，34.44%的学生更偏爱去第二食堂，26.16%的学生偏爱去第五食堂，第一、三、四食堂在学生的偏爱程度中属于一般水平，而只有 0.46%的学生在教师食堂就餐。

根据图 2，图 3，图 4 三图分析，学生对食堂的偏爱程度前三的食堂是：

早餐：第二食堂 > 第五食堂 > 第一食堂

午餐：第二食堂 > 第五食堂 > 第四食堂

晚餐：第二食堂 > 第五食堂 > 第四食堂

而学生用餐次数少的食堂（以用餐次数是否超过 10%为分界点）分别有：

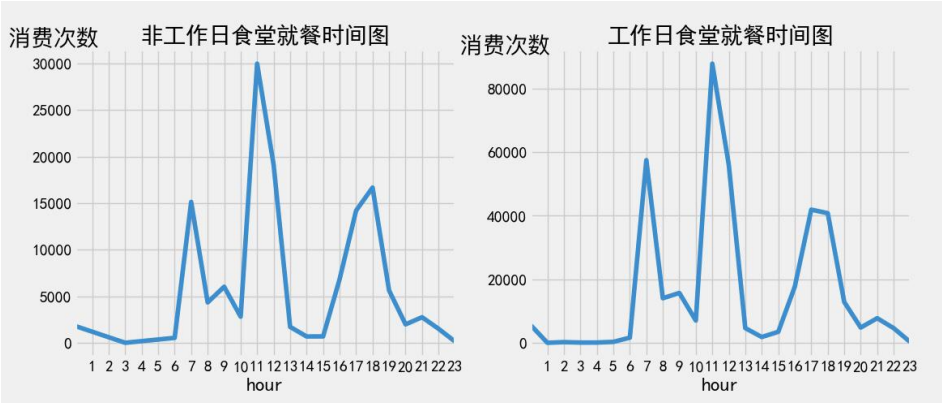
早餐：第四食堂、第三食堂、教师食堂

午餐：教师食堂

晚餐：教师食堂

综上，学生午晚两餐的用餐地点与综合三餐用餐地点分析比较，不存在显著差别；而学生的早餐用餐地点，选择第三、四食堂的占极少数，与综合三餐用餐地点有较为显著的差别。

任务 2.2 通过食堂刷卡记录，分别绘制工作日和非工作日食堂就餐时间曲线图，分析食堂早中晚餐的就餐峰值



从上图可以看出，工作日的就餐峰值均高于非工作日。工作日食堂早餐的就餐峰值为60000次，非工作日为15000次；工作日食堂午餐的就餐峰值为90000次，非工作日为30000次；工作日食堂晚餐的就餐峰值为17000次，非工作日为41000次。

出现该现象的主要原因在于工作日学生需要外出上课，直接前往食堂就餐的可能性更高，而非工作日学生由于直接在宿舍点外卖或者外出游玩就餐等原因导致前往食堂就餐的人数大幅减少。因此工作日食堂就餐峰值高于非工作日就餐峰值。

任务 2.3 根据上述分析的结果，为食堂的运营提供建议。

学校方面，应该根据学生的喜好程度合理安排食堂的场地、资金分配等资源，由2.1可知，大部分学生偏爱去第二食堂和第五食堂，因此学校应给予第二食堂和第五食堂资源倾斜。

食堂方面，受偏爱的第二食堂和第五食堂应该进行菜品创新，形成顾客粘性。并且因为就餐学生多，食堂更应该合理安排食堂内的排队位置，提高排队效率。而就餐学生数偏少的第一、三、四食堂应该找出自身原因，采取例如提高食堂环境质量、增加菜品种类或提出促销活动等方法吸引学生群体。

此外，每个食堂在就餐峰值（分别为7点、11点、17点左右）应加大食堂人手，合理安排排队场所，提高排队效率，避免打饭效率低下，并且应在这三个高峰时间段内增加菜品供应量，避免供不应求。而在非高峰期，食堂可以适当减少菜品供应和食堂工作人员数量，从而减少食堂无用的运营成本。

任务 3 学生消费行为分析

任务 3.1 根据学生的整体校园消费数据，计算，并选择 3 个专业，分析不同专业间不同性别学生群体的消费特点。

3.3.1 本月人均刷卡频次和人均消费额

根据程序计算结果得出：本月人均消费频次为：72.74118014361537次

本月人均消费额为：288.7773899469248元

考虑数据合理性，得出：本月人均消费频次越为：73次；本月人均消费额288.8元

3.3.2 选择 3 个专业，分析不同专业间不同性别学生群体的消费特点

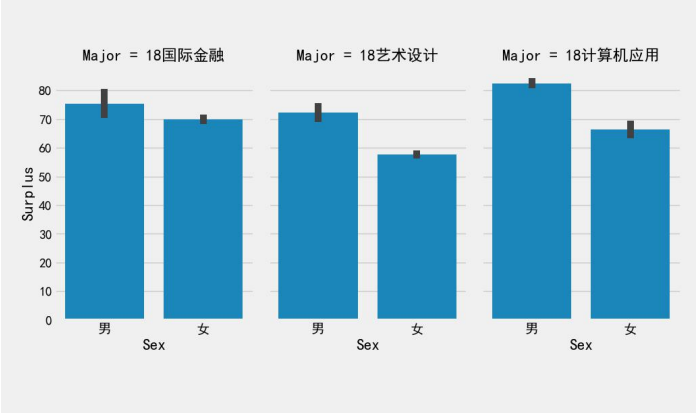
3.3.2.1 根据程序运行结果得出学生消费总额、消费次数总数、校园卡中余额的数据特征图

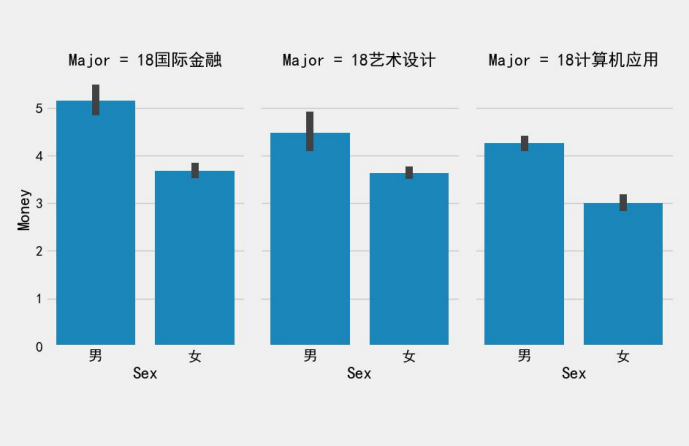
18艺术设计消费特征数据表				
	CardNo	Money	Surplus	CardCount
count	4116	4116	4116	4116
mean	183560.697	3.902186589	62.29849611	607.0515063
std	33.41934744	5.215145681	47.18377063	253.4361098
min	183493	0.1	0	77
25%	183532	1.5	27.1	401
50%	183572	3	53.05	594
75%	183589	5.5	84.9	786.25
max	183610	120	248.6	1193

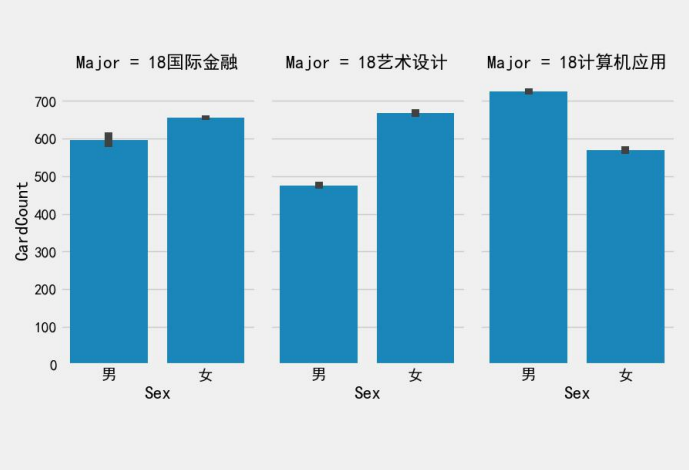
18计算机应用消费特征数据表				
	CardNo	Money	Surplus	CardCount
count	4409	4409	4409	4409
mean	183075.3019	3.973717396	78.89076435	689.2347471
std	506.3941415	4.705979982	53.69759359	243.920533
min	182829	0.2	0	127
25%	182849	1	36.1	534
50%	182866	3	70.3	675
75%	182888	6	109.3	857
max	184282	170	261.9	1222

18国际金融消费特征数据表				
	CardNo	Money	Surplus	CardCount
count	4207	4207	4207	4207
mean	180612.1196	3.859227478	70.56194676	649.1214642
std	1376.373051	5.308952442	49.67149987	191.9978694
min	180001	0.1	0	148
25%	180017	1	33.6	511
50%	180045	3	62.5	657
75%	180065	6	95.375	787
max	183910	200	318.3	1065

3.3.2.2 根据程序运行结果得出学生消费总额、消费次数总数、校园卡中余额的柱状图







从上图和上表可以得到不同专业的学生，计算机应用专业学生消费最频繁，国际金融专业学生单次消费金额最高，艺术设计专业学生卡内盈余最低。而不同专业的学生卡内盈余相差不大。出现该差异的可能原因在于计算机应用专业需要运用到电脑等电子设备，导致购买频繁。国际金融专业消费金额高可能是其运用专业知识赚钱所需。艺术设计专业学生卡内盈余最低可能是由于其日常在服装等上面的开销较大。

此外，我们可以得到不同专业间不同性别学生群体的消费特点。

首先是国际金融专业的学生。该专业女生消费频繁，男生单次消费金额高，卡内盈余金额近似。其次是艺术设计专业的学生。该专业女生消费频繁，男生单次消费金额高。男生卡

内盈余金额高于女生。最后是计算机应用专业的学生。该专业男生消费频繁、单次消费金额高，并且男生盈余金额高于女生。

通过分析，出现性别上消费特点差异主要是由于男女性格原因。女生更偏好高频低费用的购买，享受消费的过程，因此消费次数多，每次都只是购买小额商品。而男生更偏好于低频高费用的购买，消费目的性强，虽不经常消费，但每次总是会消费较大额度。

任务 3.2 根据学生的整体校园消费行为，选择合适的特征，构建聚类模型， 分析每一类学生群体的消费特点。

3.2.1 概述

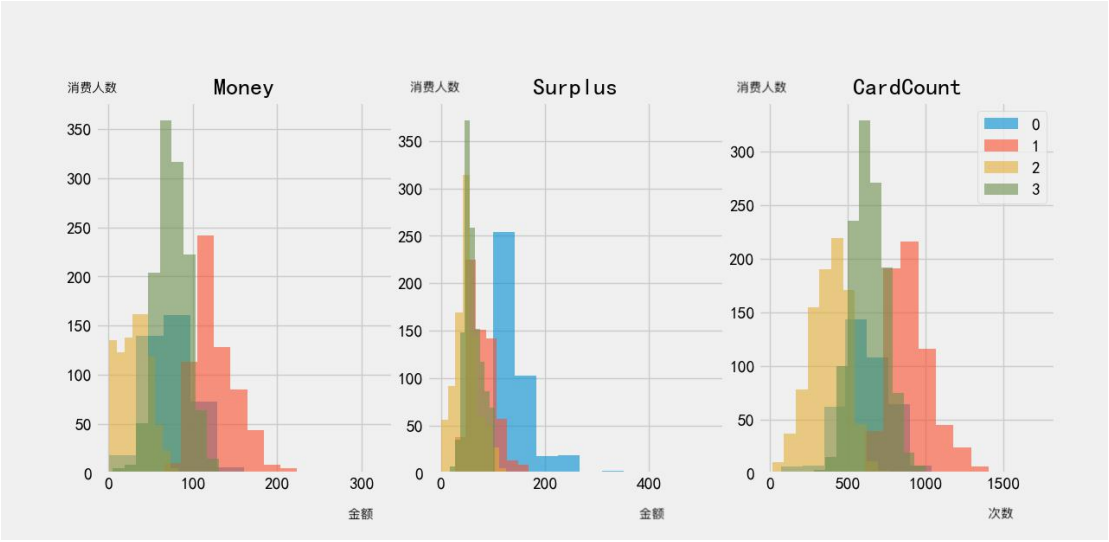
为了将学生的整体校园消费行为进行分类，选择了当月消费总金额，消费次数，卡内存款作为特征进行聚类，采用的聚类算法为 k-means 算法（k-均值聚类算法）

3.2.2 k-means 算法简介

k-means 算法（k-均值聚类算法）是一种基本的已知聚类类别数的划分算法。它是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。它可以处理大数据集，且高效。它的输入自然是数据集和类别数。聚类结果是划分为 k 类的 k 个数据集。

3.2.3 过程

将学生的整体校园消费行为分为 4 类，因此将 k-means 算法中的 k 值取为 4，运用公式 data = 1.0*(data - data.mean())/data.std() 进行数据标准化，采用欧式距离作为度量，并画出每一项特征对应的数据直方图如下



3.2.4 聚类结果分析

根据学生在 4 月份的消费金额、卡内盈余与消费次数，我们将学生分成了四类群体，分别命名为 0，1，2，3.

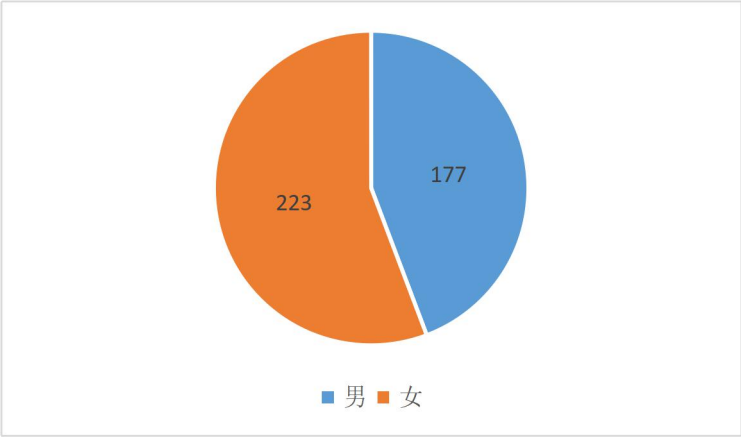
学生群体 0 的消费特点为：该群体属于中等消费水平，有较高的消费潜力，这类学生群体应有较为良好的储蓄意识，属于滞后消费。

学生群体 1 的消费特点为：该群体属于高消费水平，但消费潜力较弱，这类学生群体的消费能力较高。

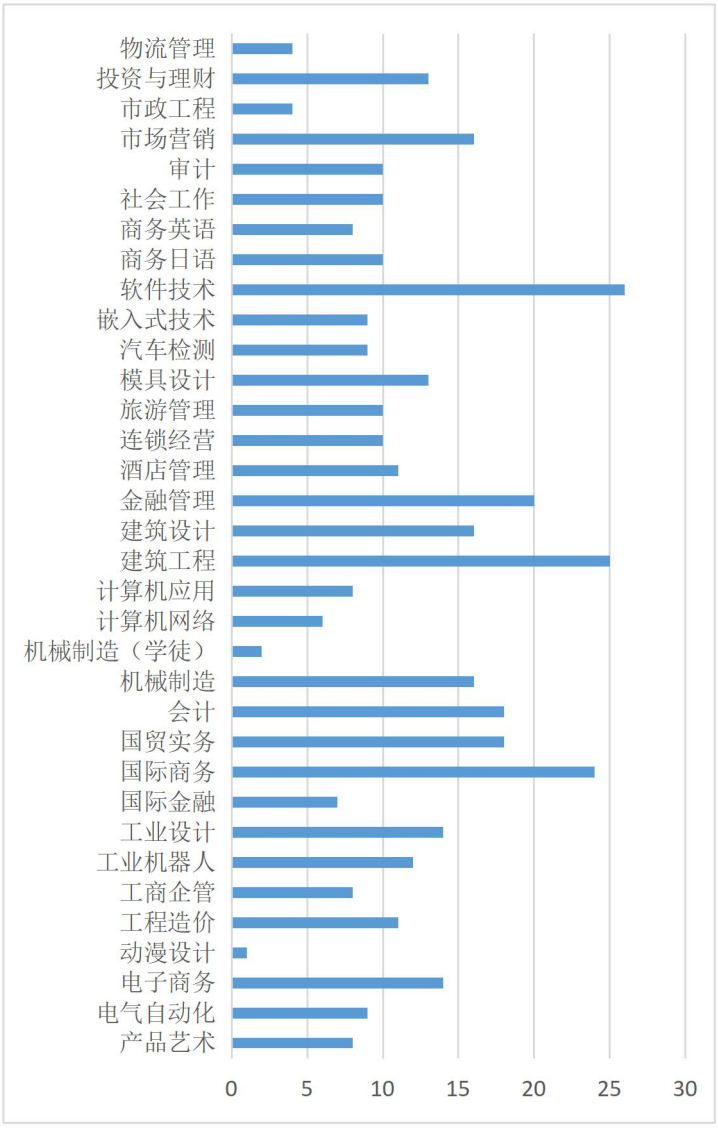
学生群体 2 的消费特点为：该群体属于低消费水平，且消费潜力较弱，这类学生群体的消费能力较弱。

学生群体 3 的消费特点为：该群体属于中等消费水平，消费潜力较弱，这类学生群体的储蓄意识较于学生群体 0 更弱。

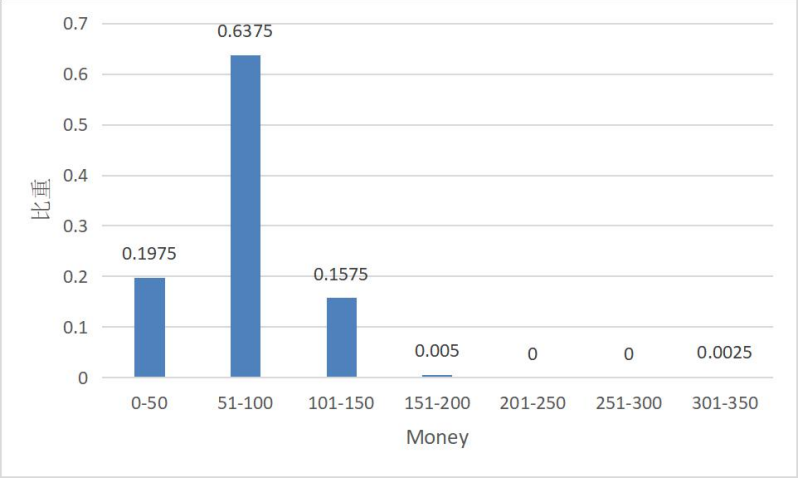
任务 3.3 通过对低消费学生群体的行为进行分析，探讨是否存在某些特征， 能为学校助学金评定提供参考。

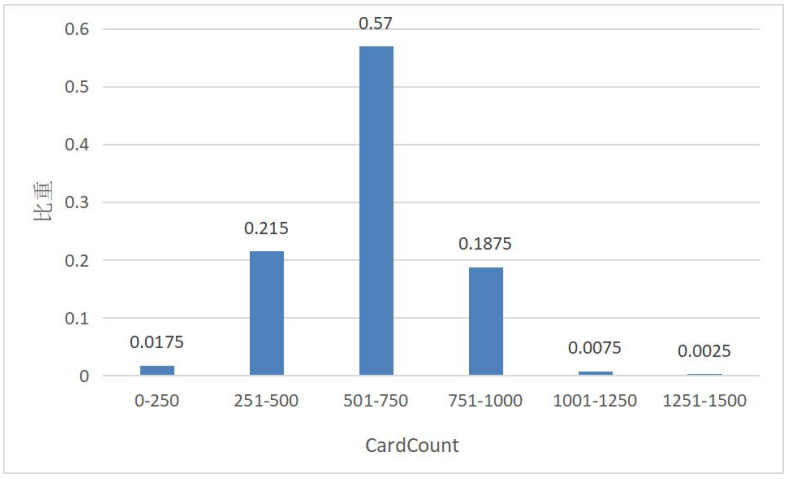


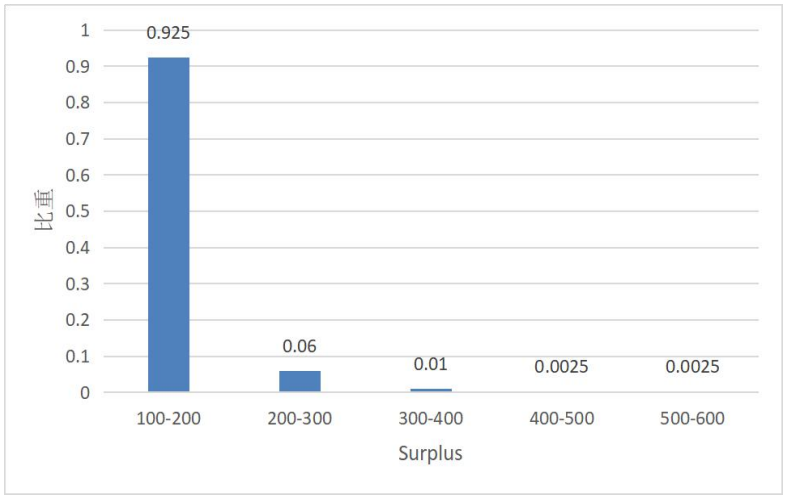
根据分类所得的贫困生情况，我们首先针对贫困生的性别进行分析。由上图可看出，在已知的贫困生人数之中，超过半数的贫困生为女性。



根据分类所得的贫困生情况，对贫困生所在专业类别进行分析归纳。可发现，专业为理工科的学生中，贫困人口占比大；经管商科的学生中，贫困人口的占比数相较于理工科的会更少。而专业为艺术设计类的学生，贫困人口数量最少。







从上图可以看出贫困生的单次消费金额主要在 51-100 元之间，消费次数主要在 501-750 元之间，卡内盈余主要在 100-200 元之间。和其他类别学生相比，我们可以看出贫困生的消费次数、消费金额和卡内盈余均较低。

从上述分析我们可以看出，贫困人口有较大概率集中在性别为女，专业为理工科，日常消费次数、消费金额以及卡内盈余都较低的学生当中。因此，学校在评定奖助学金的过程中，可以根据学生的性别、专业和日常消费情况对学生的贫困背景进行一个初步的估计，为后面对学生群体贫困背景的详细调查，提供一个简单的基础。