# 一、背景介绍

## 1、项目背景

校园一卡通是集身份认证、金融消费、数据共享等多项功能于一体的信息集成系统。在为师生提供优质、高效信息化服务的同时，系统自身也积累了大量的历史记录，其中蕴含着学生的消费行为以及学校食堂等各部门的运行状况等信息。因此，对校园一卡通数据进行分析，有利于改进学校服务并为相关部门的决策提供信息支持。

## 2、分析目标

**项目目标**

根据国内某高校校园一卡通系统一个月的运行数据，使用数据分析和建模的方法，挖掘数据中所蕴含的信息，分析学生在校园内的学习生活行为，为改进学校服务并为相关部门的决策提供信息支持。

**任务目标**

任务1 数据探索与预处理

任务1.1：根据实际项目需求对数据进行必要的数据探索和数据预处理。

任务1.2：对学生个人信息表和消费记录表进行关联。

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。

任务2.2：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。

任务2.3：根据上述分析的结果，为食堂的运营提供建议。

任务3 学生消费行为分析

任务3.1：根据18级学生的整体校园消费数据，计算本月人均刷卡频次和人均消费，分析不同专业间不同性别学生群体的消费特点。

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，为学校判定学生的经济状况提供参考意见。

# 二、数据说明

## 1、数据来源

某学校2019年4月1日至4月30日的校园一卡通数据和学生个人信息数据。

## 2、数据量

校园一卡通数据：519367 rows ×12 columns

学生个人信息数据：4341 rows × 5 columns

# 三、数据预处理

**任务1.1：**根据实际项目需求对数据进行必要的数据探索和数据预处理。

数据扩展名为csv用read\_csv导入数据，数据探索用head()展示前几行数据，shape显示数据大小，describe()查看数值分布情况。

对于学生信息表，数据包含的字段有：Index ，CardNo，Sex，Major，AccessCardNo即序号，校园卡号，性别，专业，门禁卡号。通过describe()查看数值分布情况发现，表中没有缺失值，除了性别，专业，其他均为int型数据。对CardNo做重复值检测，没有重复值

对于消费记录表，数据包含的字段有，Index，CardNo，PeoNo，Date，Money，FundMoney，Surplus，CardCount，Type，TermSerNo，conOperNo，Dept，即序号，校园卡号，校园卡编号，消费时间，消费金额，存储金额，余额，消费次数，消费类型，消费项目序列号，消费操作的编码，消费地点。

对表做数值分布查看，其中TermSerNo，conOperNo存在大量缺失值，因此对这两个字段进行删除。

对表做重复值处理，发现处理前与处理后表的大小一致，说明表没有重复值。

对表做异常值处理，学生一般不会在半夜进行消费，所以要剔除半夜消费的数据。把时间列转为时间序列，提取小时作为新列，提取时间段为6:00-23:00的数据。最后用to\_csv保存处理好的表。

**任务1.2：**对学生个人信息表和消费记录表进行关联。

用merge合并学生信息表，预处理后的消费记录表，连接字段为CardNo。

# 四、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。

要对食堂进行分析，所以选用预处理后的消费记录表，并提取就餐地点为食堂的数据。要分析学生的早中晚就餐地点，要先划分早中晚的时间段。将时间列转为时间序列，提取小时，天作为新的两列。提取6:00-10:00的数据作为早上就餐地点的时间段，中午的时间段是11:00-14:00，晚上的时间段是17:00-19:00。接下来要统计每个食堂早中晚的就餐人数，一般每个人在每个时间段只会就餐一次，所以分别对早中晚的数据框用groupby对字段day，Dept做分组，再用nunique()求出CardNo的人数，由此统计出早中晚每个地点的就餐人数，并绘制饼图，展示早中晚每个食堂的人数占比。早中晚绘制所得的图分别如图1，图2，图3：

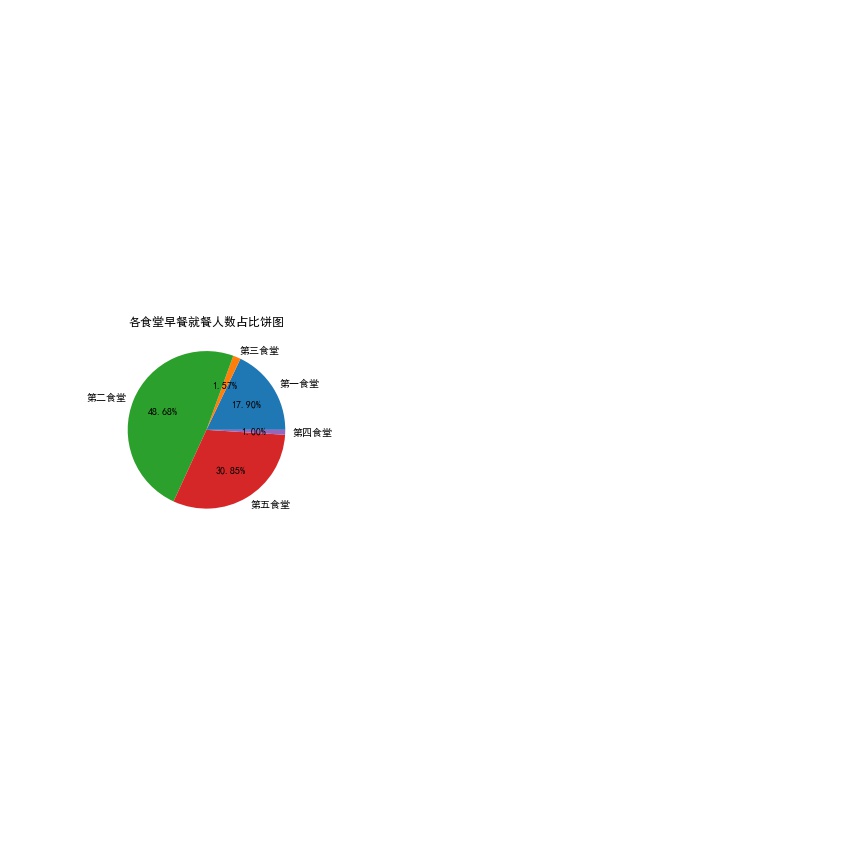


图 1

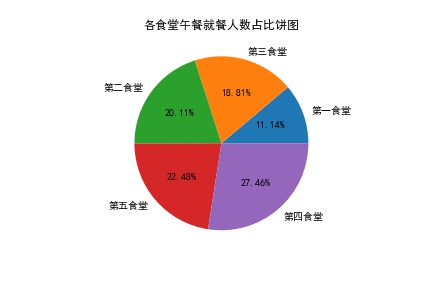


图 2

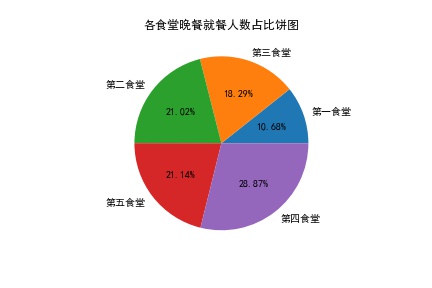


图 3

由图可知，学生早餐更倾向于到第二食堂就餐，其次是第五食堂，第一食堂，第三食堂和第四食堂早餐的就餐人数是最少的。就餐地点的人数占比差异较大。学生午餐晚餐的就餐地点与早餐的就餐地点差异较大，午餐晚餐的就餐地点基本是一样的，在人数占比上只有略微差异。在午餐晚餐的就餐地点最受欢迎的是第四食堂，其次就是第二，第五食堂了。在午餐晚餐中，第一食堂相对的就餐人数会较少。

**任务2.2：**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。

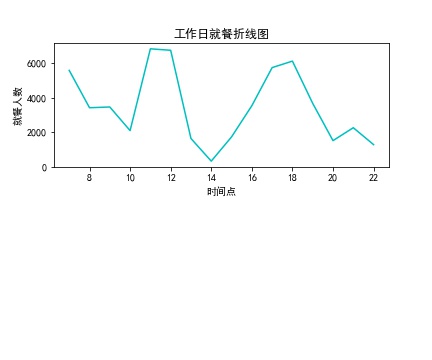
提取非工作日的数据，先按星期提取周六周日的数据，2019年4月5日是周五，是清明节假日，也是非工作日，在已经提取数据的基础上提取day等于5的数据，4月28日周日，劳动节调休，属于工作日，应在原来的基础上提取日期不为28的。同理，可提取工作日的数据。按小时分组聚合，用nunique()统计CardNo的就餐人数，作为y轴的值，x轴的值去对应的小时，绘制工作日食堂就餐时间图，如图4，非工作日食堂就餐时间图，如图5

图 4

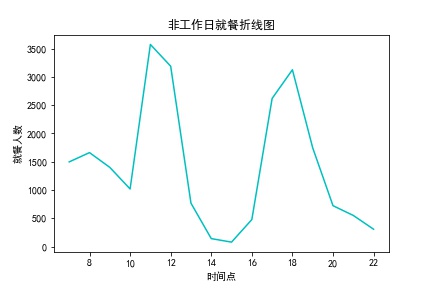


图 5

由图可知，工作日的就餐时间峰值在早上在8点之前，中午在11:00-12:00间，晚上在17:00-18:00间，工作日是学生要上课，下课，符合实际。非工作日的就餐峰值早餐是在8:00，11:00，18:00，峰值的集中趋势更明显。

**任务2.3：**根据上述分析的结果，为食堂的运营提供建议。

通过以上分析，在食堂开放方面建议早餐不开放第三食堂，第四食堂，降低运营成本，第一食堂的就餐人数相对较低，可进一步分析，对其进行整改，吸引更多学生就餐。非工作日的就餐人数少于工作日，食堂可食材的采买准备。在8:00，11:00，18:00的就餐峰值时间段可多设置窗口。

# 五、学生消费行为分析

**任务3.1：**根据18级学生的整体校园消费数据，计算本月人均刷卡频次和人均消费，分析不同专业间不同性别学生群体的消费特点。

学生个人信息表是18级的学生，合并后的消费记录表即18级学生整体校园消费数据表，选取该表进行分析统计。对Type提取类型是消费的数据，对Type字段count()计算出刷卡总次数，用nunique()求出CardNo即18级的学生人数，由此可求出人均刷卡频次为69次，对Money字段求出总的消费金额，结合18级的人数可得出人均消费为278.73元。对不同专业的分析，先提取专业的数据，这里选取了18国际金融，18会计，18连锁经营，18机械制造，18产品艺术作为代表进行分析。对Major，Sex分组聚合，统计各个专业里各性别的人数，刷卡数量，消费金额，用agg函数聚合求出。再相应的计算各个专业里各性别群体的人均消费。据此可以绘制不同专业不同性别的人均消费簇状柱形图，x轴是专业名称，y轴是人均消费金额，得到如图6：

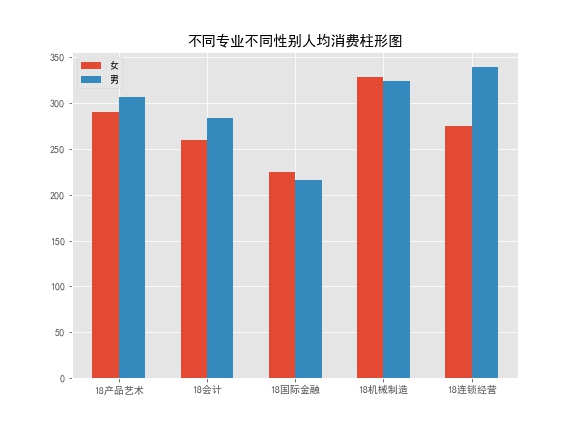


图 6

由图分析可知，大多数专业男生的消费金额要比女生的消费金额高，少数专业男生的消费金额比女生的消费金额低，但相差不大，消费水平持平。18连锁经营的男生消费偏高，相比其他专业，该专业男生与女生的消费水平相差较大。18机械制造和18连锁经营的学生消费水平较高，基本在人均300以上，普遍专业的人均消费是在200-300之间。18国际金融总体消费水平较低。

**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，为学校判定学生的经济状况提供参考意见。

使用预处理后的学生消费记录表进行分析，提取Type等于消费的数据，对此做数据分析。需要分析消费人群在不同消费场景的情况，然后据其进行聚类，除了考虑食堂的消费外，也要考虑其它主要场所产生的消费，这里有好利来食品店，红太阳超市。其他地点由于刷卡出现的次数较少，因此为了避免这些变量对聚类效果的影响，直接对其不予考虑。统计每个人在每个地点的刷卡频次及平均消费，用count对Type统计每个CardNo的刷卡次数，对每个学生计算他们的平均消费，将分组聚合后的索引转为列，把消费地点数字化。选用基于划分的 K-Means 聚类，选取的聚类特征是：每个学生的刷卡频次，人均消费，消费地点。用拐点法求出k合适的取值，得到如图7：

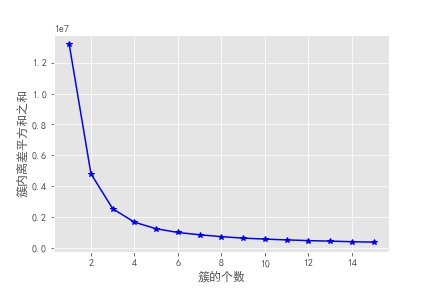


图 7

如图，随着簇数k的增加，簇内离差平方和的总和在不断减小，当k在3附近时，折线的斜率的变动就不是很大，所以选取的k值是3，即聚成3类。用采用均值-方差标准化的方法对数据做标准化处理，再建立K-Means聚类模型，得到的聚类模型图如下：

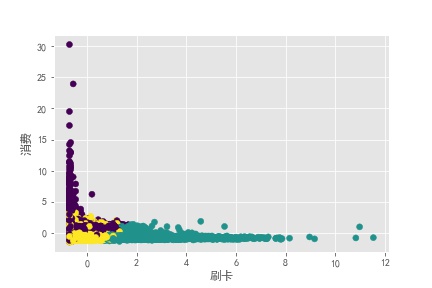


图 8

有图可知，可将模型分为3类：第一类刷卡频次低，消费水平也低，是低消费人群，相对应的可能经济条件是较低的一类人群；第二类刷卡频次高，消费水平低，应该属于喜欢消费，但经济条件受限导致每次消费只能低消费，经济条件中等；第三类是刷卡频次低，但消费高的一类，这一类应属于经济条件较好的人群。