**目录**

**[一、 背景介绍](#_bookmark0)** [2](#_bookmark0)

[1、 项目背景 2](#_bookmark1)

[2、 项目目标 2](#_bookmark2)

**[二、 数据说明](#_bookmark4)** [3](#_bookmark4)

[1、 数据来源 3](#_bookmark5)

[2、 数据采集 3](#_bookmark6)

[3、 数据量 3](#_bookmark7)

**[三、 数据探索与预处理](#_bookmark8)** [4](#_bookmark8)

[1、 对数据进行探索与预处理 4](#_bookmark9)

[2、 对个人信息表和消费记录表进行关联 5](#_bookmark10)

**四、 食堂就餐行为分析** 6

1、 绘制各食堂就餐人次占比饼图并进行分析 6

2、 绘制食堂就餐时间曲线图并进行分析 8

3、 根据分析结果，为食堂运营提供建议 11

**五、 学生消费行为分析** 12

1、 计算本月人均刷卡频次和人均消费 12

2、 分析不同专业间不同性别群体的消费特点 12

3、 选择合适特征，构建聚类模型 13

4、 分析每一类学生群体的消费特点，提供意见 14

**六、 小结** 14

# 背景介绍

## **1、 项目背景**

消费是人类通过消费品满足自身欲望的经济行为，此类经济行为是当前社会最为普遍的生活行为，大学生最为特殊的消费群体由于其独特的教育经历和所处的校园环境，使得他们的消费行为具有自己的特点。

在当前的学校生活之中,由于学生的在校时间加长,多是住在学校,日常的生活都是在学校,所以有方便的学校生活工具是极其重要的。而且在现在的学校情况下,学生的人数一直呈现增长的情况,所以有一个统一的工具来应对日益庞大的学生数量,是学校生活有序进行的关键。校园一卡通系统是有效的解决学校运转的流畅度的工具,极大的方便了校园的生活,它对应的通过学生信息的录入,相当于学生在学校的身份证,可以进行充值,然后在学校的消费活动都可以刷卡进行。校园一卡通成功的实现了学校对于信息的合理配置以及充分的运用。

此项目研究，采用数据挖掘的模式，探求学生在校园内经济行为的特点及不同自然特征学生的消费差异性，为学校建设基础设施、制定对学生消费观念的引导方向和判断学生经济状况提供依据。

## **项目目标**

1)分析学生的消费行为和食堂的运营状况，为食堂运营提供建议。

2)构建学生消费细化模型，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**

数据来源：某学校一卡通数据中心的后台数据库

数据时间：2019 年 4 月 1 日 —— 2019 年 4 月 30 日

## **2、 数据采集**

data1.csv、data2.csv对应学生信息表、消费记录表。他们的数据说明如下。

表1 data1.csv字段说明

|  |  |
| --- | --- |
| 字段名 | 描述 |
| Index | 序号 |
| CardNo | 校园卡号。每位学生的校园卡号都唯一 |
| Sex | 性别。分为“男”和“女” |
| Major | 专业名称 |
| AccessCardNo | 门禁卡号。每位学生的门禁卡号都唯一 |

表2 data2.csv字段说明

|  |  |
| --- | --- |
| 字段名 | 描述 |
| Index | 流水号。消费的流水号 |
| CardNo | 校园卡号。每位学生的校园卡号都唯一 |
| PeoNo | 校园卡编号。每位学生的校园卡编号都唯一 |
| Date | 消费时间 |
| Money | 消费金额。单位：元 |
| FundMoney | 存储金额。单位：元 |
| Surplus | 余额。单位：元 |
| CardCount | 消费次数。累计消费的次数 |
| Type | 消费类型 |
| TermSerNo | 消费项目的序列号 |
| conOperNo | 消费操作的编码 |
| Dept | 消费地点 |

## **3、 数据量**

商品销售信息：4341条 商品分类信息：519367条

# 数据探索与预处理

单纯从数据分析和机器学习建模的角度，数据的预处理是极为重要的一个环节。对于数据预处理，包括清洗、排序、缺失值的填充、异常值的删补，平稳性检测、统计量分析、相关性分析、主成分分析、滤波处理、降维、归一化等。在此次项目数据的预处理与分析过程中，本人对数据进行缺失值处理和异常值处理。

## **1、对数据进行探索与预处理**

# （1）读取数据

在正式开始进行数据分析之前，需要先读取数据。扩展名为CSV使用read\_csv，扩展名为xlsx，使用read\_excel读取。

# （2）描述统计

为了分析数据的集中趋势、数据离散程度、数据的频数分布等，进行描述性分析，描述统计是对数据进一步分析的基础。

# （3）统计缺失值

查看一下是否存由于工作人员的疏忽，忘记而缺失或者由于数据采集器等故障等原因造成缺失的缺失值。

# （4）缺失值处理

缺失值的存在会影响我们分析的效果，所以要进行缺失值处理，常用的缺失值处理有均值填充、中位数填充和最频繁值填充、或者直接删除缺失值所在的行。

# 异常值处理

打开数据表可以看到时间有异常值，2月份有30号，明显不正确。异常处理的方式有修补或者直接删除数据。

# （6）查看数据形状、数据类型

进行了缺失值，要判断是否处理成功了，这时就可以查看数据的形状了。

# （7）修改时间列为时间序列

对于时间字段，我们可以直接提取时间的年份、月份、天数、时、分等数据，但是读取进来的时间为字符串类型，所以要转化为时间序列才可以进行时间处理。

# （8）提取时间序列

Python程序能用很多方式处理日期和时间，转换日期格式、提取时间序列、计算时间差等是一个常见的功能。

# （9）数据合并

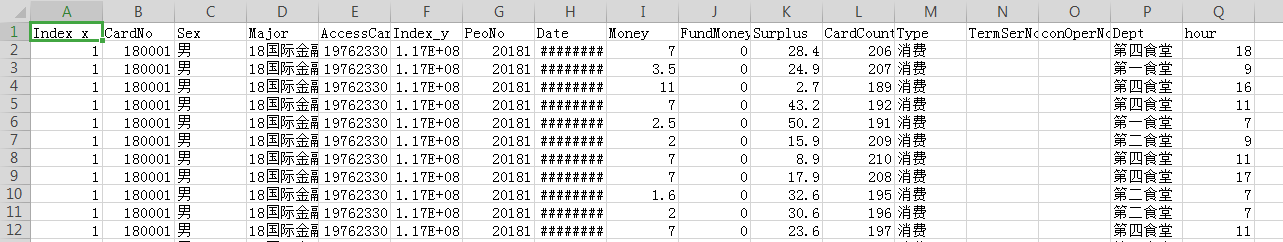
可以使用concat函数或使用merge函数合并。

# （10)保存数据

处理好一份数据，要记得保存下来。

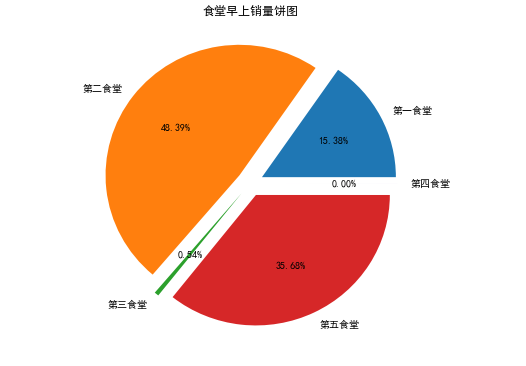
## **2、对个人信息表和消费记录表进行关联**

使用了merge函数对个人信息表和消费记录表进行关联，结果如下：

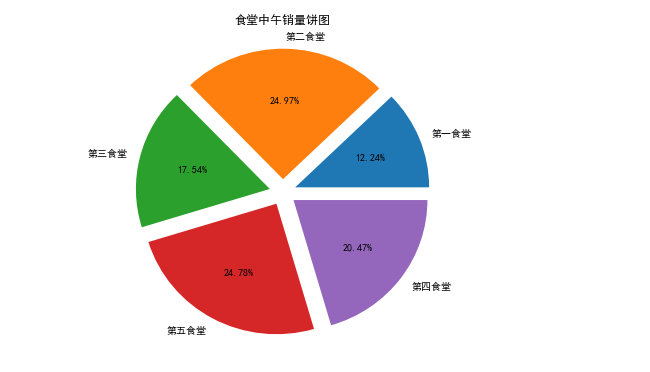


# 四、食堂就餐行为分析

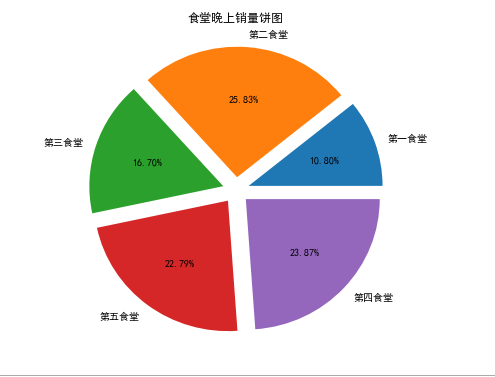
## **1、绘制各食堂就餐人次占比饼图并进行分析**



分析：从上图可以看出，有48.39%的学生早上在第二食堂就餐，35.68%的学生在第五食堂就餐，15.38%的学生在第一食堂就餐，有少数在第三食堂就餐，没有人在第四食堂就餐。这说明学生宿舍可能集中在第二食堂或第五食堂附近，而第四食堂可能没开放早餐供应。

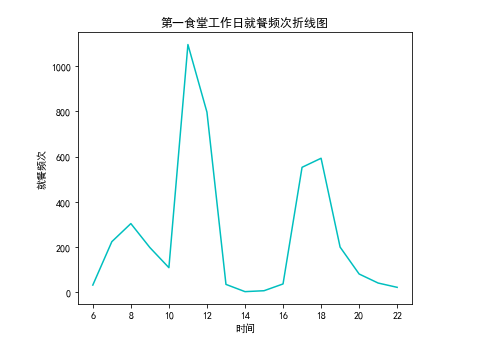


分析：从上图可以看出，有12.24%的学生中午在第一食堂就餐，24.97%的学生在第二食堂就餐，17.54%的学生在第三食堂就餐，有20.47%的学生在第四食堂就餐，24.78%的学生在第五食堂就餐。这说明学生中午就餐的地点分布相当，去第一食堂的人比较少点而已。

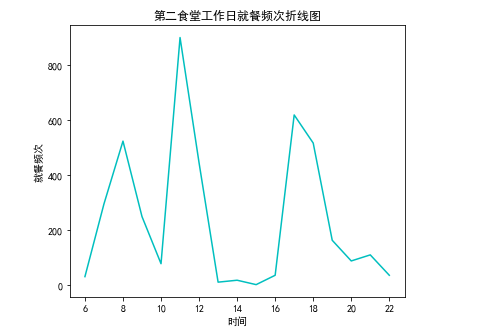


分析：从上图可以看出，有10.80%的学生中午在第一食堂就餐，25.83%的学生在第二食堂就餐，16.70%的学生在第三食堂就餐，有23.87%的学生在第四食堂就餐，22.79%的学生在第五食堂就餐。这说明学生晚上就餐的地点分布与中午的就餐地点分布差异不大，还是去第一食堂就餐的人比较少点而已。

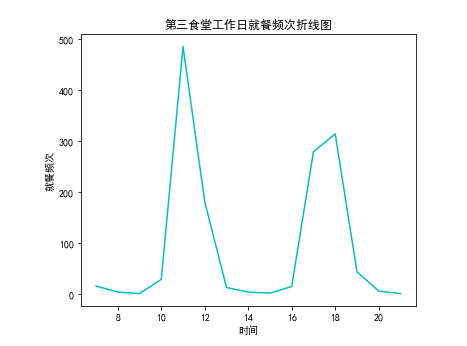
## **2、绘制食堂就餐时间曲线图并进行分析**



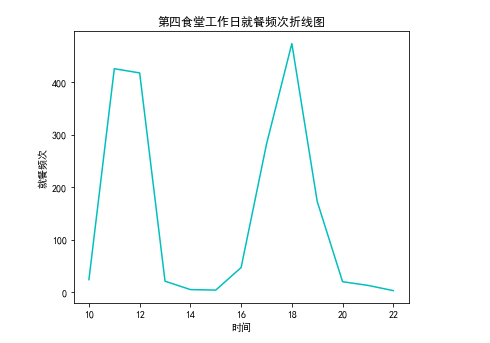
分析：由折线图看出，第一食堂工作日早上就餐峰值是早上8点；中午就餐峰值是11点30分左右；晚上就餐峰值是18点左右。



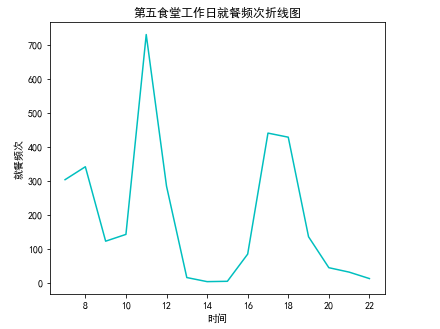
分析：由折线图看出，第二食堂工作日早上就餐峰值是早上8点左右；中午就餐峰值是11点30分左右；晚上就餐峰值是17点10分左右。



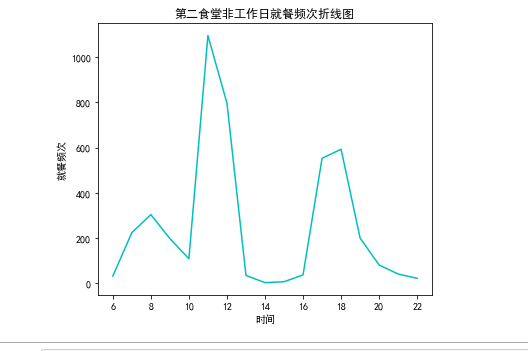
分析：由折线图看出，第三食堂工作日中午就餐峰值是11点左右；晚上就餐峰值是18点左右。



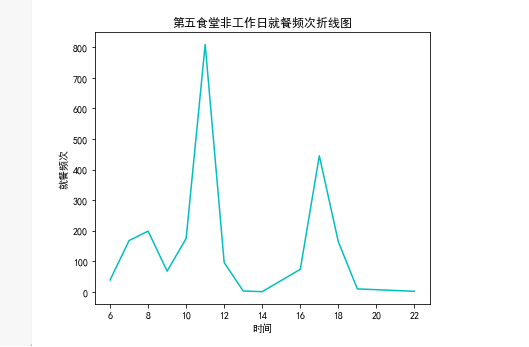
分析：由折线图看出，第四食堂工作日中午就餐峰值是11点和12点；晚上就餐峰值是18点左右。



分析：由折线图看出，第食堂工作日早上就餐峰值是早上8点左右；中午就餐峰值是11点30分左右；晚上就餐峰值是17点和18点。



分析：由折线图看出，第二食堂非工作日早上就餐峰值是早上8点左右；中午就餐峰值是11点30分左右；晚上就餐峰值是18点左右。



分析：由折线图看出，第五食堂非工作日早上就餐峰值是早上8点左右；中午就餐峰值是11点30分左右；晚上就餐峰值是17点30分左右。

## **3、根据分析结果，为食堂运营提供建议**

1、针对学生就餐地点差异，我认为第三食堂由于早餐就餐人数不多，可取消早餐供应。第二食堂可加大早餐供应量，应增加工作人员，维持好就餐秩序及保证供应速度。

2、由于中午和晚上就餐地点各食堂都分布想当，所以每个食堂都应合理分配人员和食材供应。第一食堂相对较少人用餐，可适当减少供应，以免浪费。

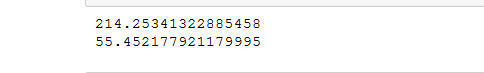
3、针对工作日早中晚就餐峰值差异，我认为食堂早餐的供应时间应为7点到9点；午餐供应时间应为11点到13点；晚餐供应时间应为17点到19点。第三和第一食堂不开设早餐供应。

4、针对非工作日早中晚就餐峰值差异，我认为开设第二食堂和第五食堂就可以了。早上就餐人数较少，可减少供应量，时间为六点半到八点半；中午就餐时间为11点到1点；晚上就餐时间为5点到7点。

5、各食堂之间可以多多沟通协调人员配置及规章制度。

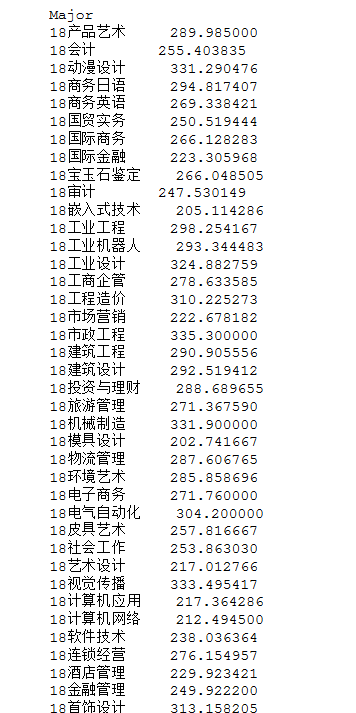
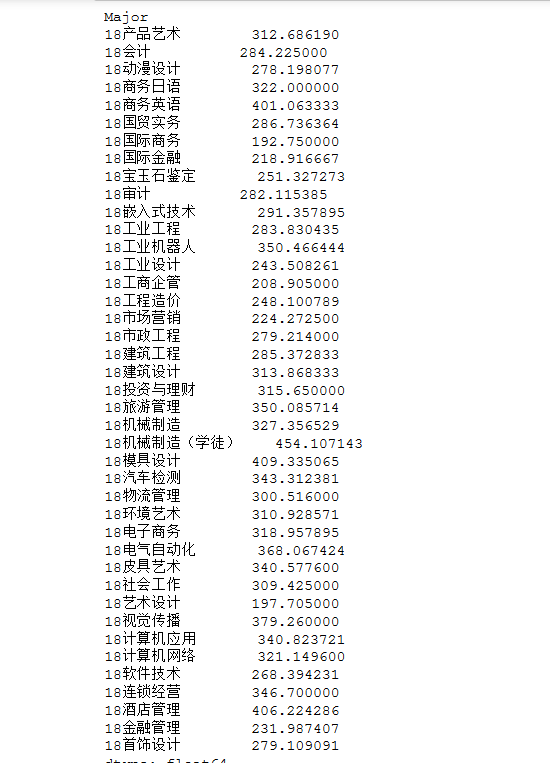
# 五、 学生消费行为分析

## **1、计算本月人均刷卡频次和人均消费**



人均消费和刷卡频次

## **2、分析不同专业间不同性别群体的消费特点**



各专业男生消费金额 各专业男生消费金额

分析：由结果可以看出，不同专业同一性别的消费金额都是有所不同的，同一专业不同性别的消费金额也是有所差异的。但总体来说，各个专业中男生人均消费均高于女生人均消费。

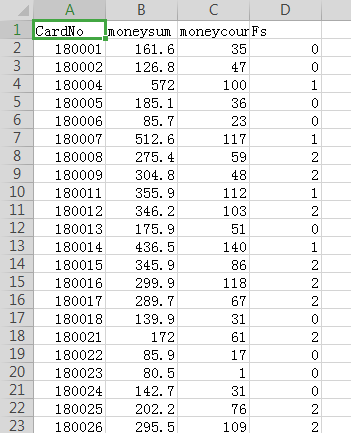
## **选择合适特征，构建聚类模型**

## （1）标准化处理，保存为task3\_2.csv。

数据的标准化（normalization）是将数据按比例缩放，使之落入一个小的特定区间。在某些比较和评价的指标处理中经常会用到，去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。其中最典型的就是数据的归一化处理，即将数据统一映射到[0,1]区间上。目前数据标准化方法有多种，常用的有均值-方差标准化、0-1标准化。

# 建立Kmeans模型：聚类分析旨在找出数据对象之间的关系，对原数据进行分组打标签，标准是每个大组之间存在一定的差异性，而组内的对象存在一定的相似性，因此大组之间差异越大，组内的对象相似度越高，最终的聚类效果就越显著。

**聚类完一部分结果如下：**



分析：从结果可以看出，共有3268条记录，分为3类，类标签分别为0、1、2。 0为校园卡消费次数较少的同学；1为校园卡消费次数较多，且消费金额较大的同学；2为校园卡消费次数较多，且消费金额较小的同学。这里需要说明的是，类标签的数值没有实际的意义，仅起到类标注的作用。

## **4、分析每一类学生群体的消费特点，提供意见**

## **（一）特点**

第一类学生群体是校园卡消费次数较少的同学，这类同学可能以点网络订餐为主，但有些餐饮地方大多证件照不齐、条件简陋、环境恶劣，存在严重的食品安全隐患。

第二类学生群体是校园卡消费次数较多且消费金额较大的同学。这类同学可能比较注重饮食健康和均衡。经济状况较好。

第三类学生群体是校园卡消费次数较多且消费金额较小的同学。这类同学可能经济状况较差。

# **对策**

1. 帮助大学生获得科学正确的营养知识；
2. 调整校内食堂运营方式、结构；
3. 针对不同类型学生制定不同的鼓励措施。如奖励每月消费次数达到一定数量的学生。

# **小结**

大学校园一般占地大，不同的餐饮消费中心相隔甚远，若不能够及时掌握就餐消费人数的情况，会导致一些食堂人数爆满，一些食堂门可罗雀，导致资源不能够有限利用。大学生的饮食消费行为又是一个复杂的活动过程，它不是孤立存在的，要受到诸多内外因素的影响。所以研究大学生校园消费行为，给出合理的建议，构建预测校园内的实时消费人数模型具有重要意义，也为学校判定学生的经济状况提供重要依据。