目录

[一、背景介绍 2](#_Toc30524)

[1.1商务背景 2](#_Toc14590)

[1.2分析目标 2](#_Toc25829)

[二、数据说明 2](#_Toc1055)

[2.1数据来源 2](#_Toc6739)

[2.2数据情况说明 2](#_Toc9728)

[2.3数据预处理与分析 3](#_Toc23113)

[三、数据分析 3](#_Toc12528)

[3.1食堂就餐行为分析 3](#_Toc9767)

[3.1.1学生早中晚餐的就餐地点分析 3](#_Toc8172)

[3.1.2食堂就餐时间分析 5](#_Toc6031)

[3.1.3食堂运营建议 7](#_Toc28228)

[3.2学生消费行为分析 7](#_Toc11255)

[3.2.1不同专业间不同性别学生群体的消费特点 7](#_Toc32085)

[四、建模分析 11](#_Toc14856)

[4.1聚类分析原理 11](#_Toc16633)

[4.2模型建立与分析 11](#_Toc18107)

# 一、背景介绍

## 1.1商务背景

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。

## 1.2分析目标

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。具体为：

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.构建学生消费细分模型，为学校判定学生的经济状况提供参考意见。

# 二、数据说明

## 2.1数据来源

数据来源：项目方提供的学生个人信息表，消费记录表

起止时间：2019年4月1日--2019年4月30日

## 2.2数据情况说明

学生个人信息表：提供了18级学生的校园卡号，性别，专业名称，门禁卡号等数据；

消费记录表：提供了学校学生的校园卡号，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，消费地点等数据。

## 2.3数据预处理与分析

学生个人信息表的Cardno列（校园卡号）和AccessCardNo列（门禁卡号）都是唯一的，消费记录表的CardNo和PeoNo列也是唯一的，两张表可以通过Cardno列进行主键连接；

消费记录表中，若出现以下情况，则为异常值：

* 消费金额，存款，余额，消费次数为负；
* 消费地点为食堂，且消费时间在夜间0-5点；

通过对数据的初步分析，进行以下预处理：

1. 对学生个人信息表的Cardno列（校园卡号）和AccessCardNo列（门禁卡号）进行去重处理；
2. 删除学生个人信息表的AccessCardNo列和消费记录表的PeoNo列；
3. 检查两张表格是否有缺失值并进行填充；
4. 检查两张表格是否有异常值并进行删除；
5. 对学生个人信息表和消费记录表通过Cardno列进行主键连接关联。

# 三、数据分析

## 3.1食堂就餐行为分析

##### 3.1.1学生早中晚餐的就餐地点分析

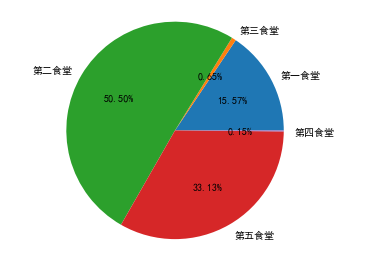


图1 各食堂早餐就餐人次占比饼图

由图1可知，该校学生早餐就餐地点中，第二食堂的就餐人次占比最高，为50.50%，大致占了1/2；第五食堂的就餐人次占比次之，为33.13%，接近1/3；第一食堂的就餐人次占比为15.57%；而第三食堂和第四食堂的就餐人次占比则比较小，其中第四食堂的就餐人次占比最低，为0.15%；第三食堂的就餐人次占比为0.65%。

该校学生早餐大多在第二食堂和第五食堂就餐，极小部分在第三食堂和第四食堂就餐。

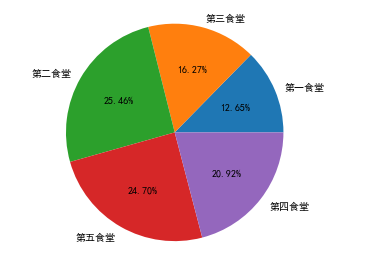


图2 各食堂午餐就餐人次占比饼图

由图2可知，该校学生午餐就餐地点中，第二食堂，第五食堂，第四食堂的就餐人次占比较为接近。其中，第二食堂的就餐人次占比最高，为25.46%，第五食堂的就餐人次为24.70%，略低于第二食堂；第四食堂的就餐人次占比为20.92%；而第三食堂的就餐人次占比为16.27%，第一食堂的就餐人次占比最低，为12.65%。

该校学生午餐就餐地点选择较为多样，大约有70%的学生选择在第二食堂，第五食堂，第四食堂就餐，且选择在第三食堂和第一食堂就餐的人数也不少，占比之和接近0.3。

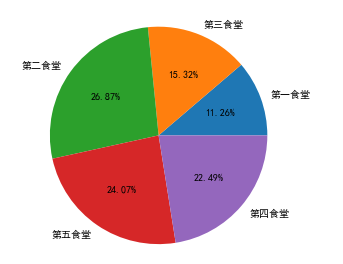


图3 各食堂晚餐就餐人次占比饼图

由图3可知，该校学生晚餐就餐地点中，第二食堂，第五食堂，第四食堂的就餐人次占比较为接近，分别为26.87%，24.07%，22.49%，其中，第二食堂的就餐人次占比最高。此外，第三食堂的就餐人次占比为15.32%，第一食堂的就餐人次占比最低，为11.26%。

该校学生晚餐就餐地点选择较为多样，选择在第二食堂，第五食堂，第四食堂就餐的人次占比接近75%，且选择在第三食堂和第一食堂就餐的人次总比之和略大于25%。

##### **3.1.2食堂就餐时间分析**

为便于分析，记就餐时间在6：00-6：59的时间为6，以此类推其他时间；

记就餐时间在6--9的为早餐就餐时间，就餐时间在11--13的为午餐就餐时间，就餐时间在16-19的为晚餐就餐时间。

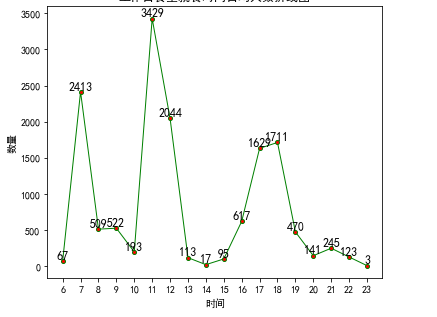


图4 工作日食堂就餐时间日均人数折线图

由4可知，工作日食堂早餐就餐高峰期主要在7：00-7：59，日均就餐人数为2413；午餐就餐高峰期主要在11：00-11：59，日均就餐人数为 3429；且在12：00-12：59为2044；晚餐就餐高峰期主要在17：00-18：59，其中，在18：00-18：59达到峰值，日均就餐人数为1711；在17：00-17：59就餐人数则略次之，为1629。

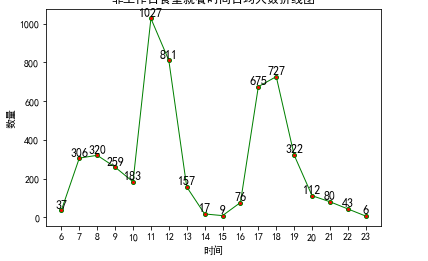


图5 非工作日食堂就餐时间日均时间折线图

由5可知，非工作日食堂早餐就餐高峰期主要在7：00-9：59，在8：00-8：59日均就餐人数为320；在7：00-7：59的日均就餐人数次之，为306；午餐就餐高峰期主要在11：00-11：59，为1027；且在12：00-12：59为811；晚餐就餐高峰期主要在17：00-18：59，其中，在18：00-18：59达到峰值，日均就餐人数为727；在17：00-17：59则略次之，为675。

##### **3.1.3食堂运营建议**

综上可知，该校学生早餐就餐地点主要在第二食堂和第五食堂，极少数学生选择在第三，第四食堂就餐；午餐和晚餐的就餐地点主要在第二食堂、第四食堂和第五食堂，且午餐和晚餐在各食堂就餐人数占比变化不大，比较稳定。此外，工作日食堂早餐就餐高峰期为7：00-7：59，非工作日食堂早餐就餐高峰期为7：00-8：59；工作日和非工作日食堂午餐就餐高峰期均为11：00-12：59；晚餐就餐高峰期均为17：00-18：59。

因此第二食堂和第五食堂可在早餐高峰时段（工作日7：00-7：59，非工作日7：00-8：59）提供较多菜品，适当增加工作人员数量；而第三，第四食堂则可适当减少甚至不提供菜品。午餐高峰时段（11：00-12：59）和晚餐高峰时段（17：00-18：59）第二食堂、第四食堂和第五食堂可提供较为丰富的菜品，并且增加工作人员数量，以更好地提供服务；同时由图4，图5看到工作日在各个时间段到食堂就餐的人数均多于非工作日，因此在工作日提供的菜品数量，种类以及工作人员数量应比非工作日有所增加。

## 3.2学生消费行为分析

##### **3.2.1不同专业间不同性别学生群体的消费特点**

经计算得出，18级学生的本月人均刷卡频次为72，人均消费为285.39。

人均刷卡频次分析

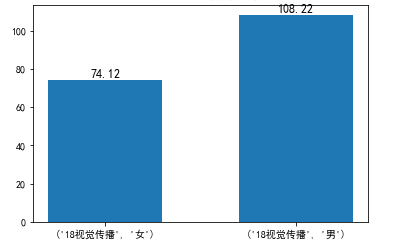
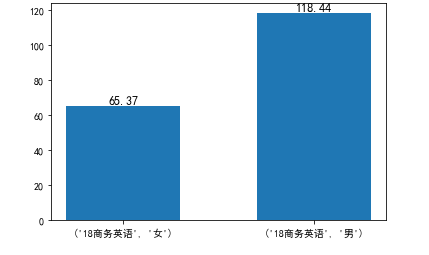


图6 商务英语专业男女人均刷卡频次柱形图 图7 视觉传播专业男女人均刷卡频次柱形图

商务英语专业的男生的人均刷卡频次最高，为118次，视觉传播的男生则次之，为102次，两者均超出18级学生的整体平均水平；而商务英语专业的女生的人均刷卡频次比较低，为65次，低于18级学生的整体平均水平；视觉传播的女生的人均刷卡频次为74次；略高于18级学生的整体平均水平。

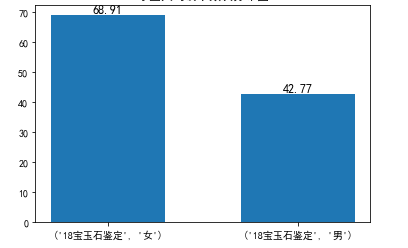
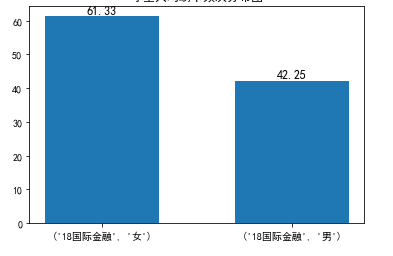


图8 国际金融专业男女人均刷卡频次柱形图 图9 宝玉石鉴定专业男女人均刷卡频次柱形图

国际金融、宝玉石鉴定专业的男生的人均刷卡频次最低，大致为42次，远低于18级学生的整体平均水平；且这两个专业的女生的人均刷卡频次也低于18级学生的整体平均水平。

人均消费金额分析

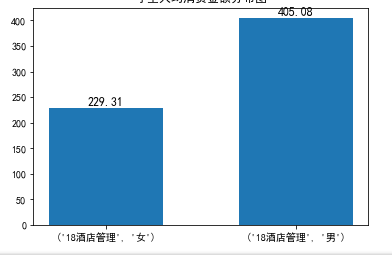
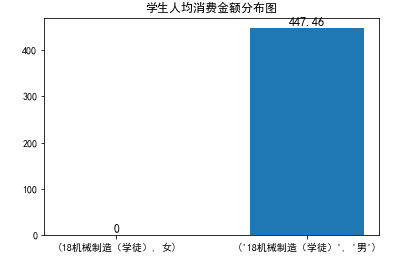


图10 机械制造（学徒）专业男女人均消费金额柱形图 图11 酒店管理专业男女人均消费金额柱形图

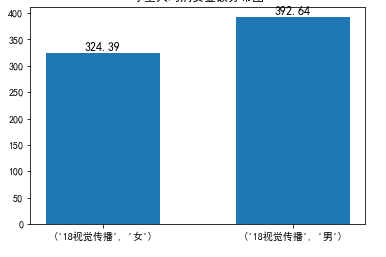
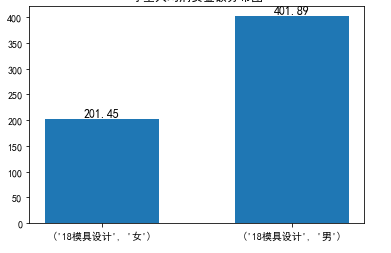


图12 模具设计专业男女人均消费金额柱形图 图13 视觉传播专业男女人均消费金额柱形图

机械制造（学徒）专业的男生的人均消费金额最高，为447.46，远高于18级学生整体的平均消费水平；由于该专业没有女生，故女生的平均消费为0；此外，酒店管理、模具设计、视觉传播专业的男生的人均消费金额也较高，分别为405.08，401.89，392.64，均高于18级学生整体的平均消费水平；但酒店管理和模具设计专业的女生的人均消费则低于18级学生整体的平均消费水平，分别为229.31，201.45；

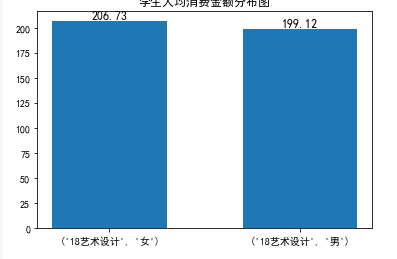
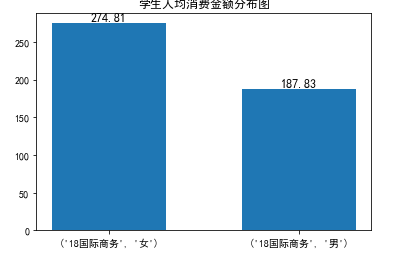


图14 国际商务专业男女人均消费金额柱形图 图15 艺术设计专业男女人均消费金额柱形图

国际商务专业的男生的人均消费金额最低，为187.83，该专业的女生的人均消费金额为274.81，略低于18级学生的整体消费水平；艺术设计专业的女生和男生的人均消费金额也较低，分别为206.73，199.12，也低于18级学生的整体消费水平。

# 四、建模分析

## 4.1聚类分析原理

聚类算法是指将一堆没有标签的数据自动划分成几类的方法，属于无监督学习方法。它是根据样本之间的距离或者说是相似性（亲疏性），把相似程度大、差异小的样本聚成一类（簇），最后形成多个簇，使同一个簇内部的样本相似度大，不同簇之间差异性高。

基本的步骤为：

step1：选定要聚类的类别数目k，随机选择k个数据点作为质心。

step2：针对每个样本点，找到距离其最近的质心，距离同一质心最近的点为一个类。

step3：聚类完成后，一共有k个集合；重新计算每个集合的质心；

step4：如果新计算出来的质心和原来的质心之间的距离小于某一个设置的阈值（表示重新计算的质心的位置变化不大，趋于稳定，或者说收敛），即可认为聚类已经达到期望的结果，算法终止。

step5：如果新质心和原质心距离变化很大，需要迭代2~4步骤。

## 4.2模型建立与分析

本文根据学生消费记录表，计算每个学生的总消费次数，平均每次消费的金额，以总消费次数（count），平均每次消费的金额（average）为指标进行聚类建模分析。

首先提取出每个学生对应的校园卡号（CardNo），计算每个学生的总消费次数（count），平均每次消费的金额（average）；其次对这两个指标进行0-1标准化预处理，使其数值在（0，1）区间内；接着对这两个指标进行kmeans聚类分析，确定最优的聚类数目。

通过轮廓系数评价法，得出下图：

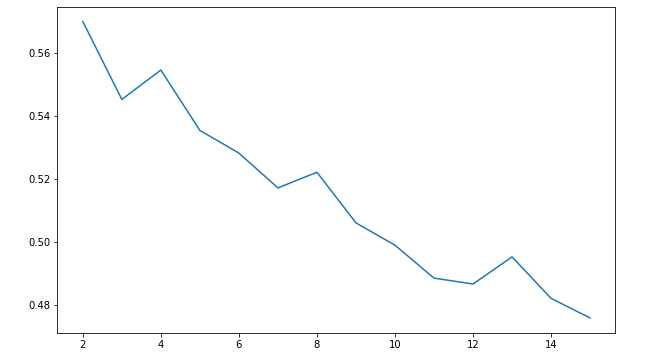


图16轮廓系数评价法结果图

由图可知，图像在2~3，4~5处畸变程度较大，因此聚类数目为3或5模型的效果较好；

通过Calinski-Harabasz指数评价法，得出聚3类calinski\_harabaz指数为：13605.678899，聚5类calinski\_harabaz指数为：16269.352465，因此综合考虑以上两种评价法，确定聚类数目为5。

以聚类数目为5进行kmeans聚类，结果如下图：

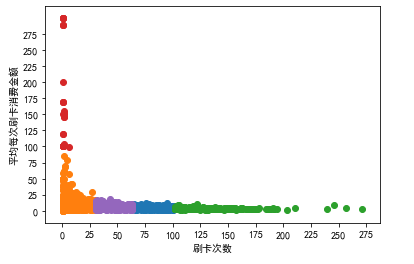


图17 kmeans聚类结果图

由图我们可把学生根据消费水平划分为5个消费群体，分别是高消费水平群体，较高消费水平群体，一般消费水平群体，较低消费水平群体，低消费水平群体；

通过分析可知，高消费水平群体人数最少，为29人，其刷卡频次在1-6之间，平均每次刷卡消费金额在98.62-300之间；

较高消费水平群体人数最多，为3000人，其刷卡频次在1-30之间，平均每次刷卡消费金额在0.1-85.3之间；

一般消费水平群体人数为2906人，其刷卡频次在31-63之间，平均每次刷卡消费金额在0.54-17.55之间；

较低消费水平群体人数为1844人，其刷卡频次在64-101之间，平均每次刷卡消费金额在1.14-11.92之间；

低消费水平群体人数较少，为641人，其刷卡频次在102-271之间，平均每次刷卡消费金额在1.48-10.23之间；

高消费水平群体、较高消费水平群体在校的刷卡次数较少，且平均每次刷卡消费金额高，这两类群体属于较为富裕的群体；一般消费水平群体在校的刷卡次数和平均每次刷卡消费金额适中，这类群体消费能力比较好；而较低消费水平群体、低消费水平群体在校的刷卡次数较多，且平均每次刷卡消费金额低，这两类群体属于贫困群体，尤其是低消费水平群体，在校的刷卡次数最多，平均每次刷卡消费金额处于低水平状态，消费能力最弱。

综上我们确定了各类消费水平群体的刷卡频次和平均每次刷卡消费金额，通过对每个学生的刷卡频次和平均每次刷卡消费金额的统计，即可确定该生的消费水平，判断其属于富裕群体还是贫困群体。