# 学生校园消费行为分析报告

**摘要：**本文以某高校“校园一卡通”2019年4月的交易流水数据作为研究对象，通过对原始数据的清洗、整合、可视化分析和聚类分析，得出结论以便帮助学校制定相关政策调整措施，做到资源的合理优化，同时也识别出“贫困生”以便校方给予一定的帮助。

**关键词：**校园一卡通、聚类分析、Python、消费行为

## 一、背景介绍

随着科技信息及网络的发展，人类进入到了信息化的时代。而高校对于数字化、信息化建设也逐步深入，“校园一卡通”系统高度集成了各类消费、小额缴费、门禁管理、会议签到管理、图书馆身份认证及借阅、学籍管理等信息。通过共同身份的认证机制，实现数据管理的分享。因此，“校园一卡通”系统很快就成为了高校信息化建设的重要组成部分。

随着时间的推移，“校园一卡通”系统中会积累大量的消费和行为数据，通过这些数据，在一定的程度上可以了解学生在校园的数字轨迹，掌握学生在校园的行为规律，对于学校的资源优化和配置有着重要的意义。

另外，“扶贫”、“脱贫”已经成为近年来的政治焦点，大学生贫困生活状况受到政府部门越来越多的关注，贫困学生的认定结果对于扶贫工作有着直接的影响。贫困资助的准确发放与贫困生的学习生活有着密切的关系，校园一卡通中的消费类数据在这里就起到了关键作用。南京理工大学教育基金会，就利用校园一卡通所获取的信息开展了“瑞华暖心饭卡”项目，该项目专门针对贫困本科生的“温饱问题”进行援助。该项目的落地就是拿全校的一万六千余名在校本科生9月中旬到11月中旬的刷卡记录，对所有的记录进行了数据分析，最终锁定了500余名“准援助对象”。

本项目以国内某高校2019年4月（共30天）的一卡通交易流水数据为研究对象，通过对原始数据的清洗、整合、可视化分析和聚类分析，得出结论以便帮助学校制定相关调整措施，做到资源合理的优化，同时也识别出“贫困生”以便校方给予一定的帮助。这里，我们首先对所有的数据进行初步探索和数据预处理；其次对食堂数据进行分析，得到各食堂就餐人次占比饼图与工作日和非工作日就餐时间曲线图；接着，根据学生整体校园消费数据和消费行为分析，得到某一专业不同性别学生群体人均消费对比柱形图；最后，我们选取了16个指标，进行聚类分析。经过聚类模型分析，为学校后勤部门提出了参考性建议，适当对食堂资源进行优化，另外，为学校的“精准扶贫”提供有效的参考依据。

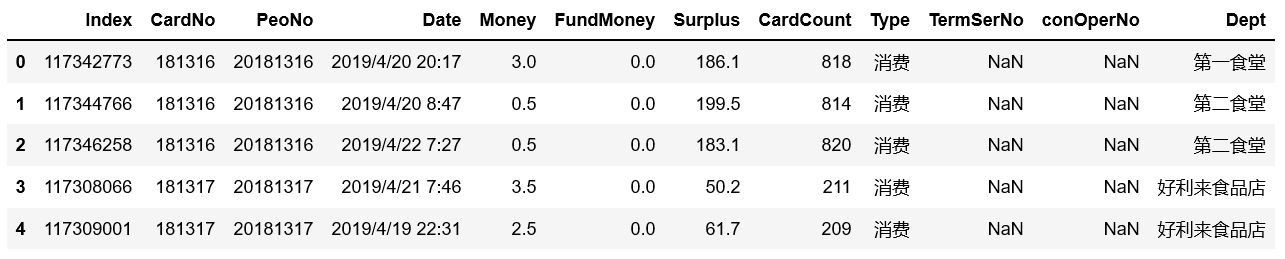
## 二、数据预处理

本项目的原始数据有两个数据表，其中一个表包含了学生基本信息，即Index、CardNo、Sex、Major、AccessCardNo这5个字段，共计4341条记录，该表的示例如图1所示。



**图1 学生基本信息表示例**

另一个表包含了学生的刷卡信息，即Index、CardNo、PeoNo、Date、Money、FundMoney、Surplus、CardCount、Type、TermSerNo、ConOpenNo、Dept这12个字段，共计519367条记录，该表的示例如图2所示。



**图2 学生刷卡记录表示例**

首先，根据实际项目需求，我们需要对数据进行必要的数据探索和数据预处理，并对学生基本信息表和学生刷卡记录表进行关联，以便未来对数据进行描述分析时可以很方便的看出不同专业不同性别学生的特点。在此，我们利用CardNo字段对两个表进行关联，并从以下4个角度对数据进行预处理：

① 删除表“data2”中的两个字段：TermSerNo、conOperNo。因为这两个字段中存在大量的缺失值。

② 通过统计“data1”中Sex的频数，检测一下该字段是否存在异常值。结果发现无异常值。

③ 查看两个数据表中是否有重复值，查看后发现没有重复值，所以没有进行重复值的处理。

④ 将“data2”的Date改为“时间类型”，并将“年份”、“月份”、“星期”、“天”、“时”、“分”、“秒”进行了提取。

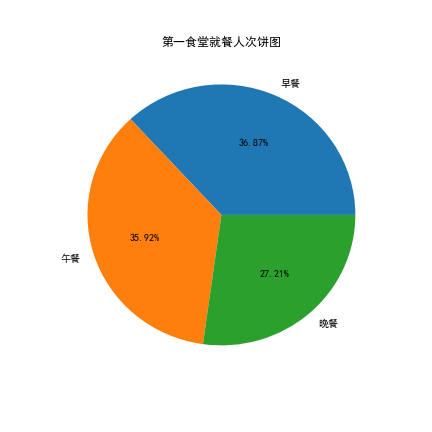
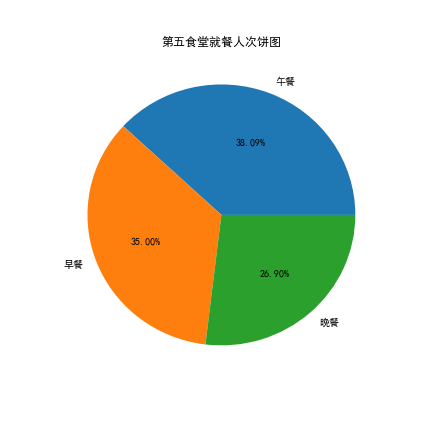
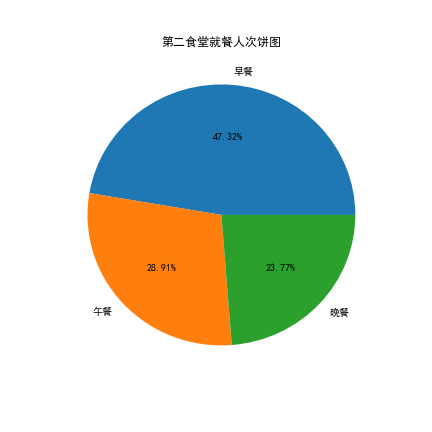
经过上述预处理，最终联合后的数据表如图3所示。



**图3 预处理后的数据**

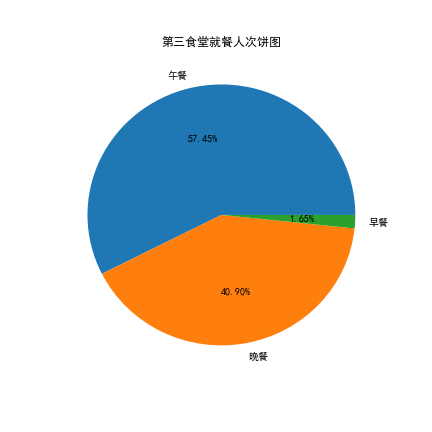
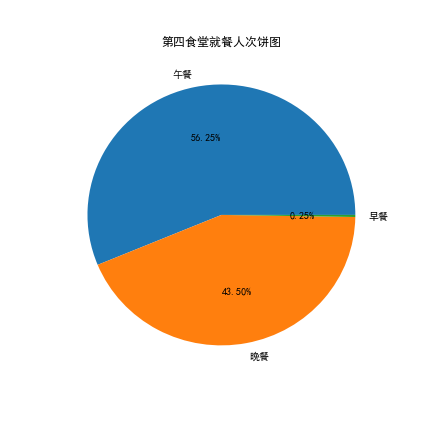
## 三、就餐行为分析

首先，我们根据人们常规的就餐习惯，将早中晚餐定义了三个区间，即6:00-9:00为早餐时间、11:00-14:00为午餐时间、17:00-20:00为晚餐时间。分别绘制出了6个饼图来表示食堂早中晚餐的就餐人数，由图形可以看出学生早中晚餐的就餐地点还是有明显区别的。

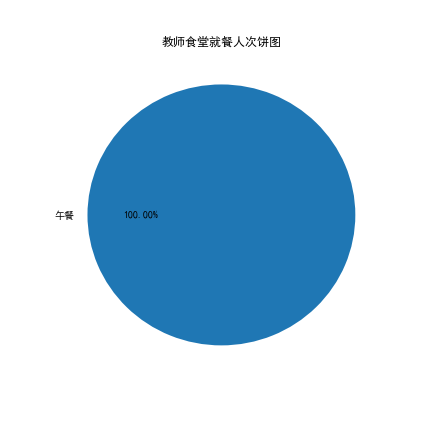
**图4(1) 第一、二、五食堂就餐人数对比图**

（1）第一食堂、第二食堂、第五食堂具有共同的特点，早中晚餐的用餐人数基本处于“平分秋色”的状态，只是第一、二食堂的早餐卖的稍好一些，而第五食堂的午餐比早晚餐略好一些。

**图4(2) 第三、四食堂就餐人数对比图**

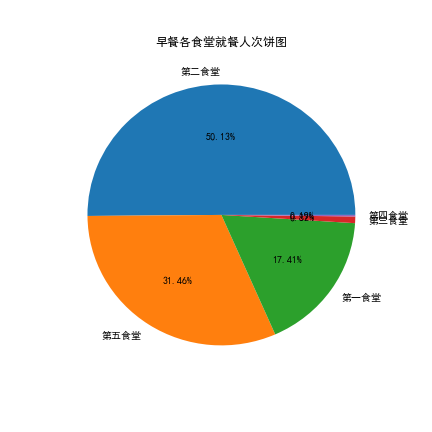
（2）第三食堂、第四食堂具有共同的特点，早餐的就餐人数极少，午餐的就餐人数最多，晚餐略少一些。



**图4(3) 教师食堂就餐人数对比图**

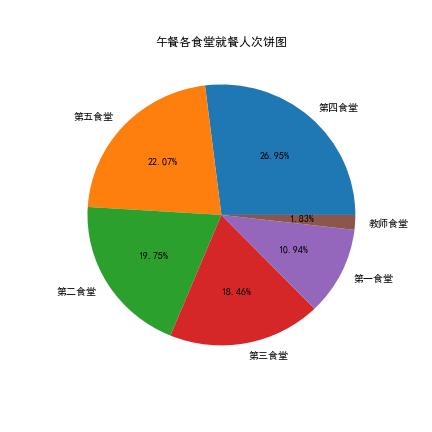
（3）教师食堂是一个特例，这个食堂只有午餐的就餐人数，换句话说就是很可能这里只提供午餐。我们猜测，由于教师早上一般吃完早饭才来上班，而中午正常情况下会在学校吃饭，晚上下班则又选择回家吃饭，所以，为了避免浪费，教师食堂只提供午餐也是合理的。

以上是从食堂角度进行的划分，我们可以看出每个食堂的早中晚餐就餐人数对比。下面，我们是从早中晚餐角度进行划分的，可以看出学生们比较偏爱哪个食堂的早中晚餐。



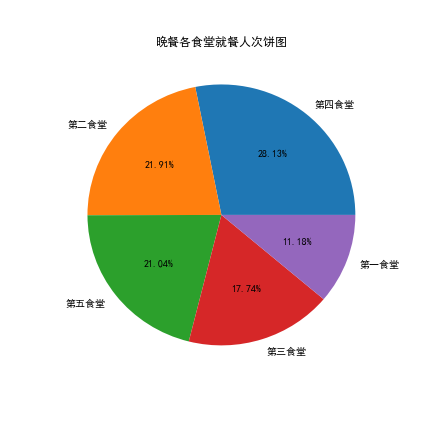
**图5(1) 早餐就餐人数对比图**

（1）早餐的生意可谓是“三分天下”。学生们对第二食堂的早餐极为偏爱，其次是第五食堂，在二、五这两个食堂就餐的总人数占到了80%+，而去第一食堂吃早餐的人数会少一些，只有17%，几乎没几个人选择去第三、第四两个食堂吃早餐。



**图5(2) 午餐就餐人数对比图**

（2）午餐的生意可谓是“平分天下”。第二、三、四、五食堂都差别不大，基本各占20%左右，其中第四食堂较为突出一些，占到了27%左右。而第一食堂稍逊一些，只有10%左右。学生们几乎不会选择去教师食堂去吃午餐。



**图5(3) 晚餐就餐人数对比图**

（3）晚餐几乎也是“平分天下”的状态，各个食堂之间差别不大，其中第四食堂最好，占28%的市场份额，第二、五食堂次之，第三食堂再次，第一食堂最差，只占11%。

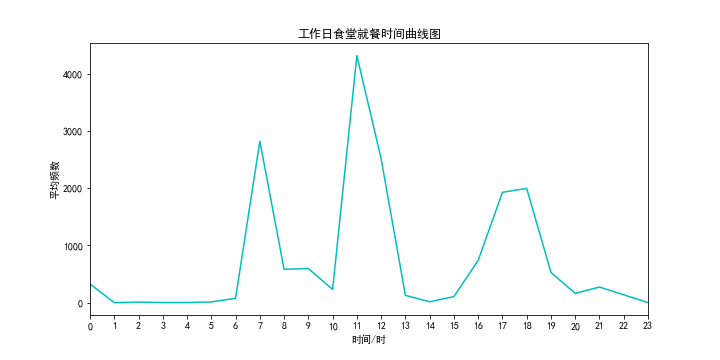
基于以上就餐行为分析，我们可以为学校的后勤集团提出如下建议：

（1）第三、四食堂只需提供午餐和晚餐即可，撤销这两个食堂早餐的售卖可以节省更多的资源。

（2）教师食堂只需提供午餐，但量也不需太大，基本够教师的消费需求即可。

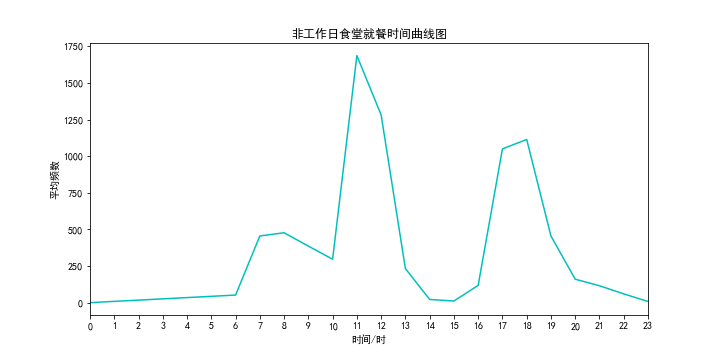
（3）第一食堂的早中晚三餐的销售都不是特别好，建议改善这个食堂的就餐环境或者增加一些新菜式，多一些促销活动，以提高其对学生们的吸引力。

（4）第二、五食堂一日三餐卖的都比较好，说明这两个食堂对学生的吸引力极大，可以加大对这两个食堂的投入力度，让它们好上加好。



**图6(1) 工作日食堂就餐时间曲线图**

由工作日的就餐时间曲线图，可以很明显的看出早中晚三个高峰期，早餐高峰为7:00左右，午餐高峰为11:30左右，这两个高峰都比较尖锐，即学生们的就餐时间点非常集中。而晚餐高峰较为平缓，持续时间较长，从17:00-18:30左右基本都是高峰时段。



**图6(2) 非工作日食堂就餐时间曲线图**

由非工作日的就餐时间曲线图，可以看出很明显的高峰有两个，即午餐和晚餐，而早餐只能算是个小高峰，而且持续时间较长，峰值较午餐和晚餐明显小很多。具体而言，早餐高峰为8:00左右，午餐高峰为11:30左右，晚餐高峰为17:00-19:00。综上所述，非工作日，由于大家会睡懒觉，所以，早餐高峰较平时推后一些，而且就餐人数低了很多，估计多数人都选择放假时不吃早餐，直接早午餐两顿合并了。午餐和工作日相比，并无明显差别。放假期间吃晚餐的人数会比平时多很多，估计是大家都比较倾向于假期搞一些聚餐活动。

基于上述分析，我们可以提出如下建议：

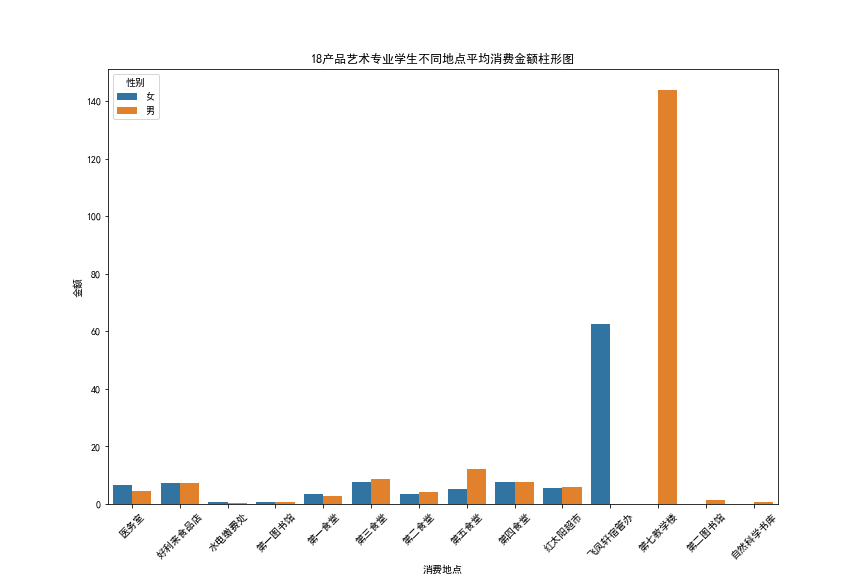
（1）各食堂应该在峰值来临前做好充分的准备，以确保在就餐高峰期快速便捷的提供服务。

（2）非工作日期间，食堂应该减少早餐的供应数量，以免浪费资源。另外，工作人员也可以享受一下假期的待遇，比以往晚一个小时开门营业。晚餐的食材应多准备一些，以便应对更多的就餐需求。

## 三、学生行为分析

**1、学生整体消费数据统计**

根据18级学生的整体校园消费数据，计算得出本月人均刷卡频次为73次/人，人均消费金额为289元/月。经过汇总统计，我们选取了“18级产品艺术专业学生在不同地点平均消费金额”为代表，来解读该专业不同性别学生群体的消费特点。



**图7 18级产品艺术专业学生不同消费地点的平均消费金额**

上图中可以看出18级产品艺术专业男女生在医务室、好利来食品店、水电缴费处、第一图书馆、第一食堂、第三食堂、第二食堂、第四食堂、红太阳超市平均消费金额相差不大，有的甚至相同；在飞凤轩宿管办只有女生消费，可以推测这里是女生公寓；而第七教学楼、第二图书馆、自然科学书库只有男生进行消费，可以推测这三个地点离男生宿舍较近，方便男生消费，或者第二图书馆和自然科学书库中有大多数男生喜欢借阅的书籍，而由于超期未还所以需要缴纳罚款。

**2、聚类模型分析**

本次任务中，我们构建了16个指标，即聚类特征。具体包括：每位学生每餐的平均消费金额（早餐平均金额、午餐平均金额、晚餐平均金额）；每位学生在不同消费地点的总消费额（教师食堂总金额、第一食堂总金额、第二食堂总金额、第三食堂总金额、第四食堂总金额、第五食堂总金额、好利来食品店总金额、红太阳超市总金额）；不同时期的平均消费金额（工作日平均金额、非工作日平均金额）；就餐时间（早餐时间中位数、午餐时间中位数、晚餐时间中位数）。对这些指标做了均值-方差标准化，并做了K均值聚类分析，根据业务需求，即从“可用来做学校判定学生经济状况的依据”角度出发，我们将18级学生分为3类。根据模型结果建议，这3类具体可命名为“高消费人群”、“中等消费人群”、“低消费人群”。由此，学校可快速识别出低消费人群，即“准贫困生”，以便出台各种帮扶措施。

## 四、结论

经过K均值聚类分析，我们得到了聚类中心表，由该表我们将得出如下结论：学生消费群体被分为3类：“0”代表低消费人群，“1”代表高消费人群，“2”代表中等消费人群。这三类人群的特点分别是：

* 低消费人群比较喜欢在第三食堂、第五食堂和好利来食品店购买食物，他们在工作日消费会高一些，非工作日也许会离开校园去做兼职工作，故非工作日的消费极低；
* 高消费人群比较喜欢在第一食堂、第二食堂、第四食堂、第五食堂以及红太阳超市消费，且他们在工作日消费会低一些，反而非工作日会高一些，初步推断，可能是由于非工作日会有些聚会或放松心情的需求，因此消费增高；
* 中等消费人群比较偏爱的第三食堂和好利来食品店购买食物，他们同样是在非工作日消费高一些。

学校可根据上述聚类分析的结果，初步锁定某些“准贫困生”，多给予一些帮扶措施。

最后，从就餐时间上看，所有人都比较偏爱在7点多吃早餐，8、9点钟的也有，但比较少，几乎没人在7点以前吃早餐，另有少部分低消费人群不吃早餐；午餐集中就餐时间一般都在11点和12点两个高峰期，很少有人在12:30以后再吃午餐了；晚餐也会有些人选择不吃，但大多数还是会选择在17-18点区间用餐，拖到19点以后的就是极少数了。

从上述结论，我们也可以探索出学生的作息习惯，从而帮助学校进行决策，对食堂的开放时间进行优化，以降低人工成本。