**学生校园消费行为分析**

摘 要

本次任务主要运用可视化展示、数据挖掘和Kmeans聚类建模的方法，探索、分析和预测近二十四万条校园一卡通运行数据。根据学生的消费时间、消费地点、消费类型、消费金额和消费次数等数据，利用Python软件分析学生在校园内的消费情况，为有关部门改进校园服务（如食堂运营）等提供有效建议和信息支持。并在此基础上选择消费金额和消费次数作为聚类特征，构建Kmeans模型，分析不同专业不同性别学生的消费特点，为学校判断学生的经济状况提供参考意见及数据支持。

**关键词：**Python，数据挖掘，消费行为，Kmeans聚类，可视化

目 录

1. 项目目标
2. 分析过程

2.1 任务一的分析过程

2.2 任务二的分析过程

2.2.1 学生就餐地点分析

2.2.2 学生就餐高峰分析

2.3 任务三的分析过程

**1、项目目标**

本次数据挖掘目标是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。

**2、分析过程**

本案例主要包括以下步骤：

1. 数据探索与预处理：在data2的原始数据上进行预处理的常规步骤，然后结合项目要求合并学生个人信息表、消费记录表。
2. 行为分析：绘制食堂就餐人次和就餐高峰的可视化图形，以此分析学生食堂就餐行为，为学校食堂运营提供建议。
3. 构建模型：根据18级学生整体校园消费数据，选择合适特征，构建聚类模型。

**2.1 任务一的分析过程**

（1）打开并读取data1、2文件。

（2）查看消费记录表的基本统计量，了解各个指标的分布情况。

（3）统计缺失值。

本次任务数据中无缺失值存在，不需采取缺失值填充方式。

（4）异常值处理。

校园卡号CardNo、消费金额Money、消费次数CardCount、消费地点Dept这四个字段已经进行了处理并未发现问题，就剩消费时间Date，打开数据表可以看到0点时食堂出现消费，明显不正常。因为食堂0点不开门，于是需采取删除操作。可以发现，只要0点食堂有消费，TermSerNo就不是空值，于是可以挑出TermSerNo列并删除带空值的行放入c中，再用~isin剔除原始数据中包含c的数据就可以删除异常值。因为TermSerNo和conOperNo这两列是无关数据，所以我在这里直接删除，不影响后续数据的运用。最后将处理的数据进行保存。

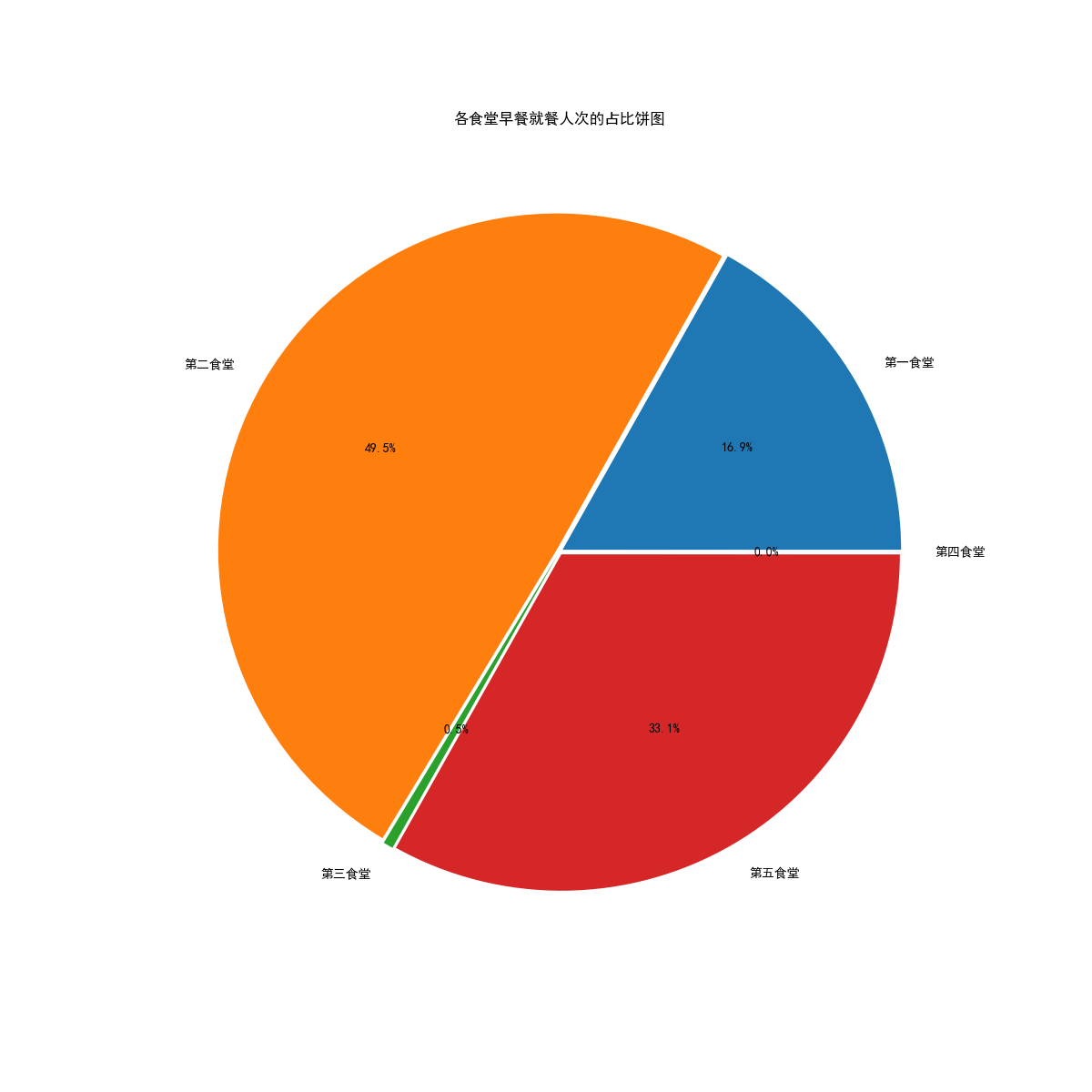
（5）关联数据。

本次案列主要分析学校18级的消费行为，因此以CardNo列为连接键用merge函数对学生个人信息表和消费记录表进行关联，得到18级学生的消费记录新表（下文用新表表示此结果）。

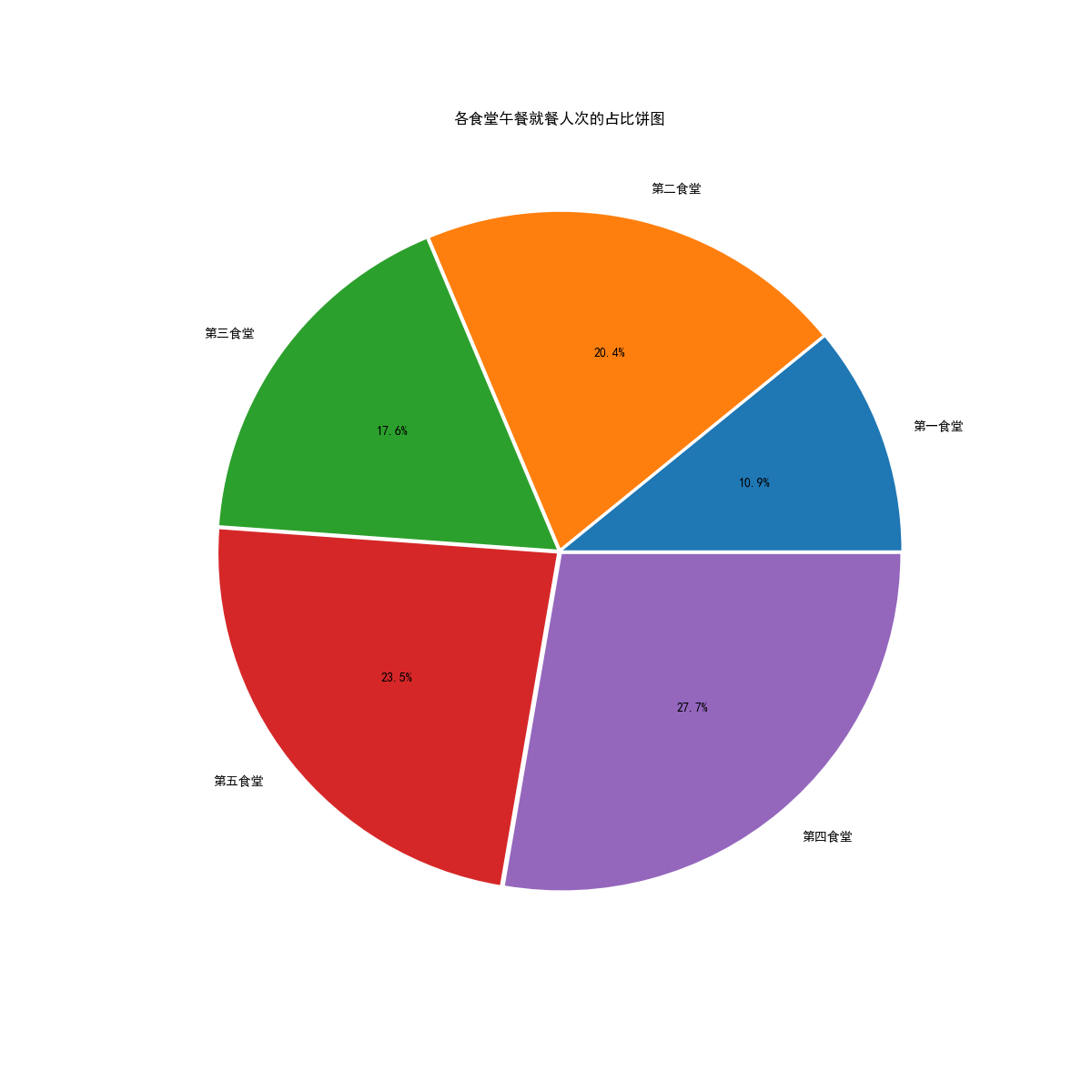
**2.2 任务二的分析过程**

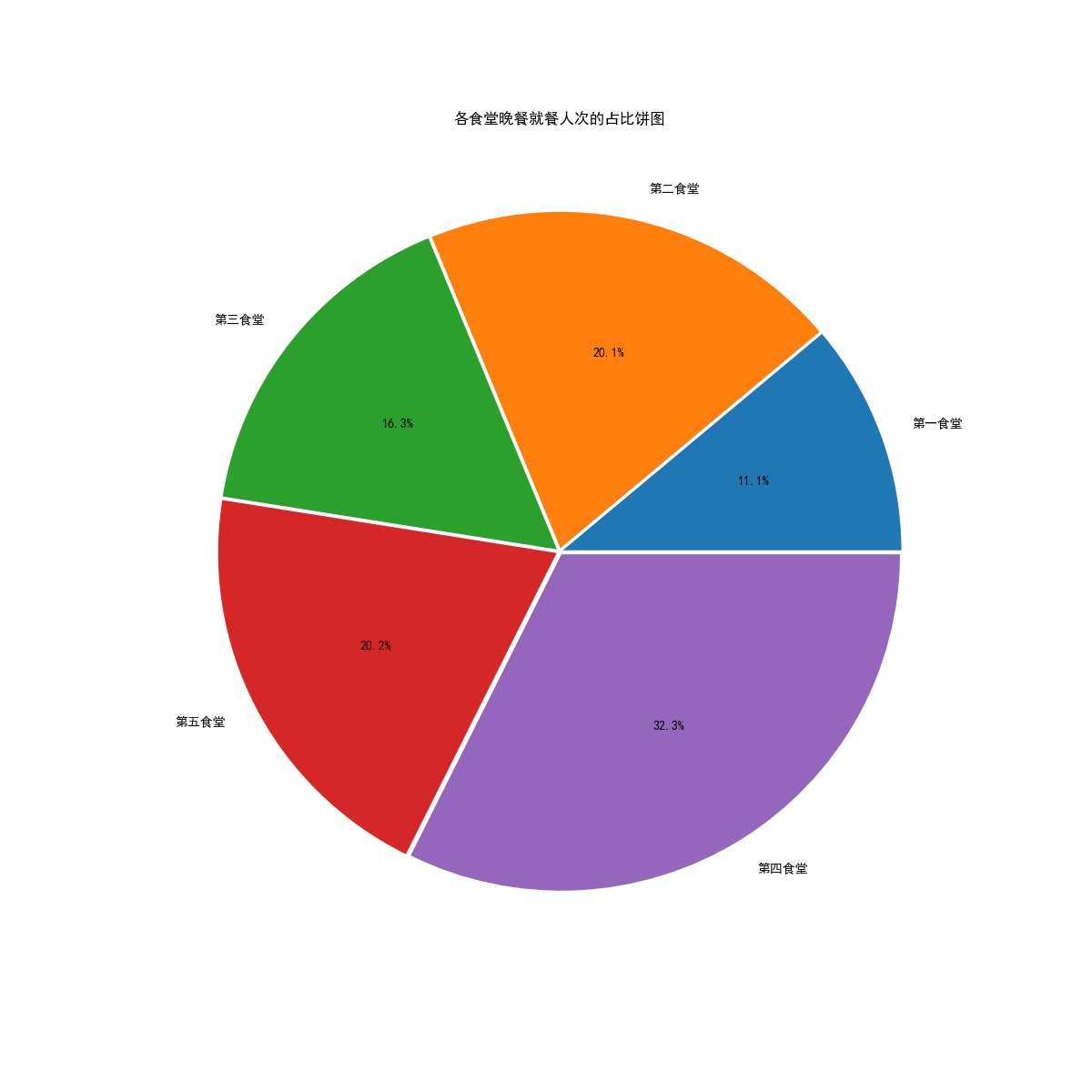
2.2.1 学生就餐地点分析

修改新表的时间列为时间序列，之后采用For循环提取消费地点为食堂的数据，再从中分别提取早中晚餐的数据（在本案列中设置6，7，8点为早餐时间，11，12点为午餐时间，18，19点为晚餐时间），然后以day和CardNo为组合计算其消费金额，得到消费记录总数就代表就餐人次，以此绘制出各食堂就餐人次的占比饼图。



从食堂早餐的就餐人次占比饼图可以看出学生就餐地点有显著差异，接近一半的18级学生选择在第二食堂就餐，第三、第四食堂最少人就餐，甚至可以忽略不记，因此第三、第四食堂早餐时间段可以不开门，第二食堂需准备更多的人手和食物数量、种类来满足学生需求。

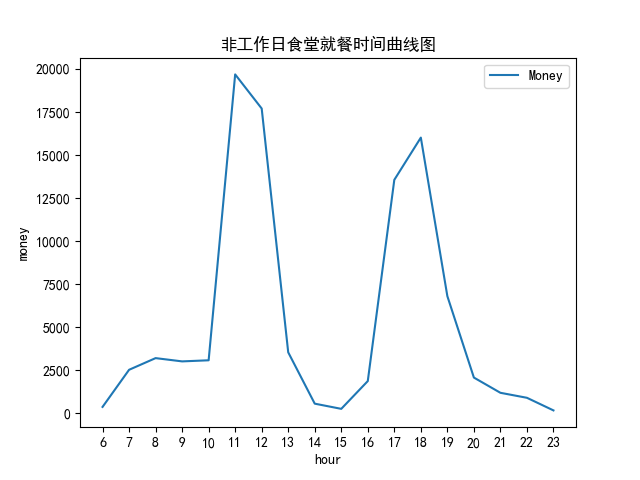


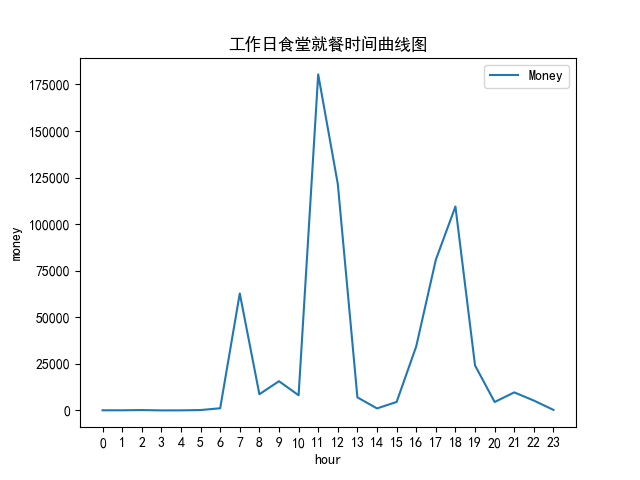


从各食堂午餐、晚餐的占比饼图可以看出的就餐人次相差不大，即学生午餐、晚餐的就餐地点无显著差异，只是相对而言第五食堂有更多的人选择。

2.2.2 学生就餐高峰分析

和2.2.1的做法相似，采用For循环提取消费地点为食堂的数据，再从中分别提取工作日和非工作日的数据，之后挑出hour和money两列，以每小时为一组计算消费金额，消费金额越多就表示该时间段就餐人次越多，即得出食堂早中晚的就餐高峰。就此绘制工作日和非工作日食堂就餐时间曲线图。





从图中可看出，非工作日的早餐高峰为8点，午餐高峰为11点，晚餐高峰为18点，工作日的早餐高峰为7点，午餐高峰为11点，晚餐高峰为18点，可以发现工作日和非工作日的就餐高峰基本一致。

根据上述可视化展示和分析结果，提供以下几条建议供学校参考：

1. 为了保障学生的生活，就餐人次较多的食堂应扩大规模，如安排更多的人手，增加食物的种类和数量等；
2. 根据就餐高峰的时间段，提前准备好食物，安排好阿姨打菜，并可适当的安排食堂工作人员协调学生就餐秩序，在安全、舒适、卫生的前提下，以高质快捷的服务保证学校师生的就餐。
3. 在食堂粘贴就餐高峰提醒，让学生错峰就餐，以此减轻食堂压力。

**2.3 任务三的分析过程**

根据18级学生的整体校园消费数据，查看新表的行数、消费金额总额和学生个人信息表的行数可以知道本月的人均刷卡频次、总的人均消费金额；在分别从新表和学生个人信息表中挑出专业、性别和消费金额三列，以专业、性别为组分别计算消费总额以及人数，将这两个数据相除可得本月不同专业不同性别学生之间的平均消费金额，以此绘制不同专业间男女的平均消费柱状图，由此分析每一类学生群体的消费特征，从而构建聚类模型，帮助学校判定学生的经济状况，对学生进行有效的援助。

构建Kmeans聚类模型前，先对数据进行均值-方差标准化，再构造聚类器。根据消费金额和消费频次，较少的一类标签设为0，中等的设为1，较高的设为2，输入新的学生数据，其标签被定义为0，则学校需要对其进行补助，若被定义为1则可适当的进行补助，定义为2则无需对其进行补助。

