学生校园消费行为分析

一·项目背景

随着中国高等教育的快速发展，高校在校学生人数逐年增加，导致管理压力剧增，很难及时发现学生经济存在的问题，影响了学生的健康发展。而校园一卡通作为在校师生的必备物品，记录了学生就餐、超市消费、其他缴费等数据，这些数据背后隐藏着大部分学生日常行为活动信息，而通过挖掘和分析学生日常行为活动可以提前了解学生的消费状况。本文利用数据挖掘技术对学生校园一卡通数据进行研究，分析了学生日常消费行为，分析和挖掘得到的结果可以帮助学校管理者及时为学业出现问题的学生提供引导和帮助以及为合理配置资源、制定科学决策提供支持和参考，促进学校的智能化管理。

二·项目目标

目标一：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

目标二：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

三·项目任务

任务一：数据预处理

1.1 根据实际项目需求对数据进行必要的预处理

1.2 分别对学生个人信息表和消费记录表，以及学生个人信息表和门禁记录表进行关联

任务二：食堂就餐行为分析

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明显差别

2.2 通过食堂刷卡记录，分别绘制工作日和非工作日食堂的就餐时间曲线图，分析食堂早中晚餐的就餐峰值

2.3 根据上述分析的结果，为食堂的运营提供建议

任务三：学生消费行为分析

3.1 根据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，并选择 3 个专业，分析不同专业间不同性别学生群体的消费特点

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

3.3 通过对低消费学生群体的行为进行分析，探讨是否存在某些特征，能为学校助学金评定提供参考

四·项目任务的解题过程

1 任务一：数据预处理与分析

1.1 任务 1 的解决

任务分析：在 data1 表中的 CardNo，AccessCardNo 都不能有重复值，故需对这两列做去重处理。在 data2 表中的时间在 0 点到 6 点之间应剔除，因为在这段时间进行消费是不合理的。

任务方法：data1 表：利用 drop_duplicates 对指定列做去重处理。

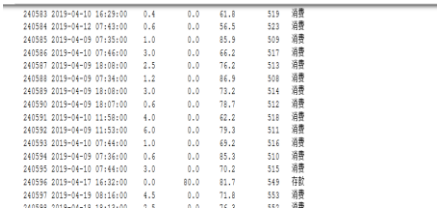
data2 表：先找出 0 点到 6 点的数据，再在 data2 表删掉这些数据。

最后将预处理后的 data1 和 data2 表根据 CardNo 为连接点，进行内连接。

空值处理：根据索引‘AccessCardNo’对 data1 数据进行去重，发现处理前原有 4341 条数据，处理后变成 4336 条数据。

统计缺失值：经过统计 data1 和 data2 的数据，发现不存在缺失值。

修改成时间序列并剔除从凌晨 0 点的数据到早上 6 点前的数据。部分结果如下：



2　任务二：数据分析与可视化

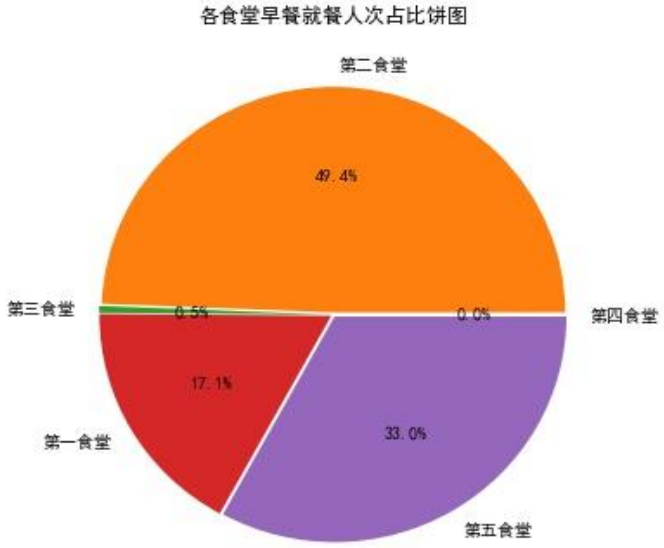
2.1　任务 2.1 的解决

绘制各食堂早上的就餐人次的饼图。

任务分析：操作过程中需要每个时间的天数和小时数，故将时间细分为天数和小时。计算第一食堂早上的就餐人次，其中在短时间连续消费也算作一人，故在同一天的 7，8 点（早餐时间定义可更改）不管消费多少次都算作一人。

任务方法：利用 pivot_table 或 groupby 函数对天数和 CardNo 二次分组，对 Money（可改为其他参数）进行 count 计数。计算第二食堂早上的就餐人次的方法也如此，故可利用 for 循环将五个食堂早上的就餐人次计算出来并画图。

各食堂早餐的就餐人次的饼图如下：



可以看出，在早上最受学生欢迎的食堂是第二食堂，销量占比总体的 48.39%，其次是第五食堂，占 35.68%，最低的是第三食堂，销量惨淡；在中午，依旧是

第二和第五食堂销量最高，从总体看，五个食堂销量差距不是很大；在晚上，第二食堂占 25%，第四食堂占 24%。由此可以分析出，面对高流量的第二食堂，可以提高第二食堂的菜品数量和丰富程度，扩大食堂的用餐空间，提供多个取餐窗口等，方便学生打饭和就餐。

各食堂午餐、晚餐就餐人次绘制方法同早餐就餐人次。

各食堂午餐的就餐人次的饼图如下：

可以看出占比相差并不会特别大，占比大小顺序为第四食堂>第五食堂>第二食堂>第三食堂>第一食堂。

各食堂晚餐的就餐人次的饼图如下：

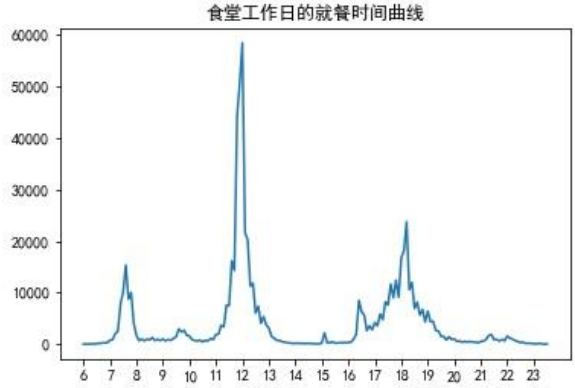
可以看出占比大小顺序为第四食堂>第二食堂>第五食堂>第三食堂>第一食堂。午餐和晚餐食堂就餐人次比例相近，其中第四食堂午餐，晚餐人数最多，深受学生喜爱，不提供早餐。第二，第五食堂一整天都有较大的人流量。第一食堂整天都只有比例较小的人流量，学生最不喜欢的食堂，第三食堂不提供早餐的供应，人流量仅高于第一食堂。

2.2 任务 2.2 的解决

任务分析：任务需要计算食堂工作日的就餐时间曲线，故先提取食堂的数据，用时间得到星期数可提取工作日（星期数为 1-5），因用小时所做曲线不够平滑，可以利用分钟数来绘制图形。

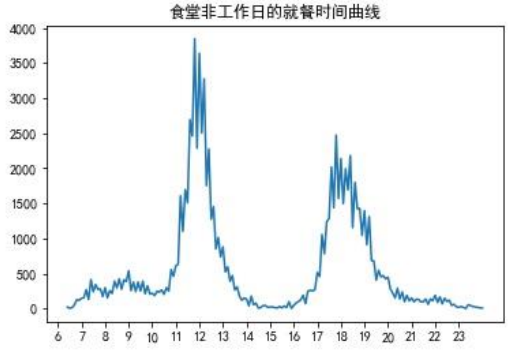
任务做法：用 groupby 对分钟分组，绘制每分钟的 Money 总数绘制折线图（消费越高，就餐人数就会越多）。

食堂工作日的就餐时间曲线绘制如下：



可以看出早上就餐高峰为 7-8 点，中午就餐高峰为 11.30-12.30，晚上就餐高峰为 5.30-6.30。

非工作日就餐时间曲线做法与工作日就餐时间曲线做法相近，绘制如下：



可以看出在工作日是 7-10 点消费的金额相近也就是就餐人数相近，原因可能是非工作日同学起床时间不统一，中午和晚上的就餐时间大体与工作日就餐时间相近。

2.3 任务 2.3 的解决

给食堂的意见：

- 适当改进食堂一，使得更多人去该食堂吃饭，增加同学们的选择。

- 可以在食堂三和食堂四增设早餐的服务。
- 在 11 点前准备好午餐，在 17 点前准备好晚餐。
- 在工作日时早上 7 点准备好早餐，早上 8 点左右就可撤去早餐服务。在非工作日时可适当延长早餐的就餐时间。

3 任务三：生成自动售货机的画像

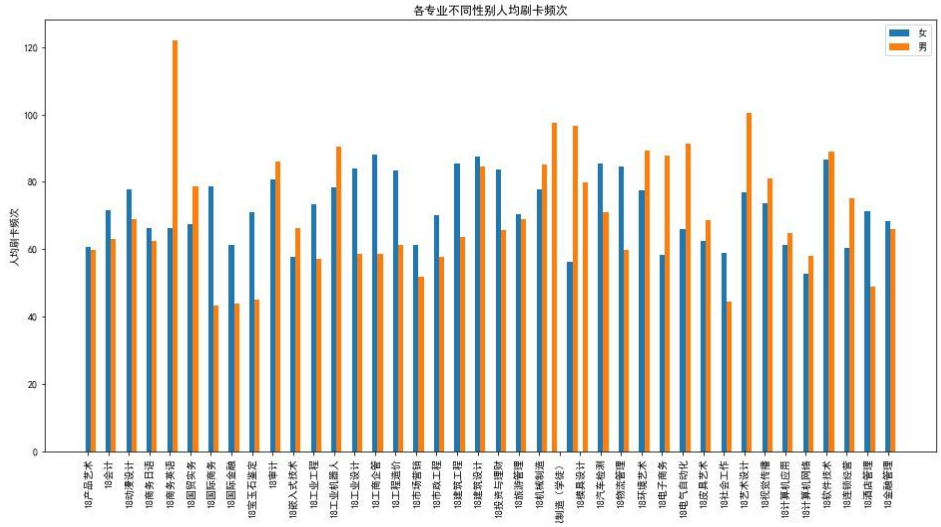
3.1 任务 3.1 的解决

任务分析：任务 3.1 可分为俩个任务，任务①为计算总体的人均刷卡频次和人均消费，任务②为分析不同专业不同性别学生群体的消费特征，可以利用不同专业不同性别学生群体的人均刷卡频次和人均消费来分析消费特征。

任务①做法：先将 CardNo 进行分组，求出每个人的刷卡次数，再加总，除以个数就得到人均刷卡频次，计算得人均刷卡频次为 72.66。人均消费也是先将 CardNo 进行分组，求出每个人的消费总数，再加总，除以个数就得到人均消费，人均消费为 281.16。

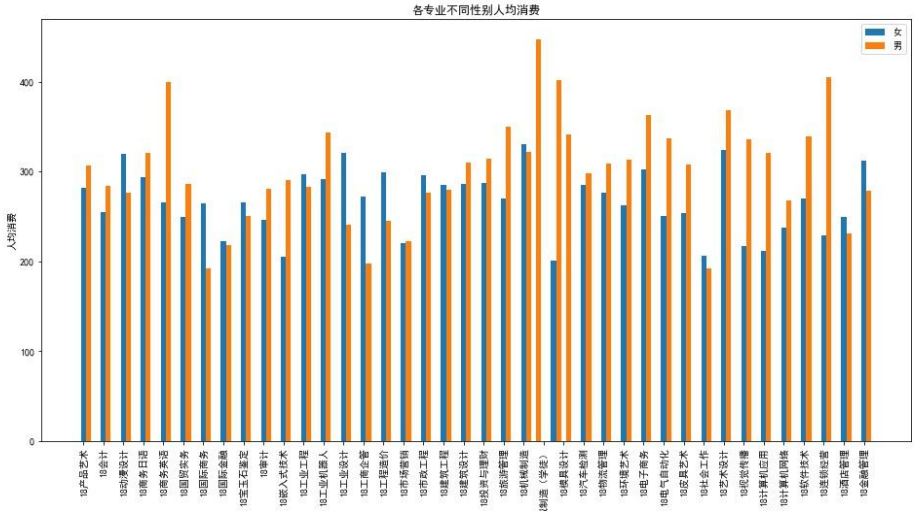
任务②做法：对 Major 和 Sex 进行分组，得到每个专业不同性别的消费总额和消费频次，再计算每个专业不同性别的人数，就可以计算不同专业不同性别学生群体的人均刷卡频次和人均消费。

不同专业不同性别学生群体的人均刷卡频次柱状图绘图如下;



可以看出大部分专业女生的刷卡频次都略大于男生的刷卡频次。

不同专业不同性别学生群体的人均消费柱状图绘图如下;

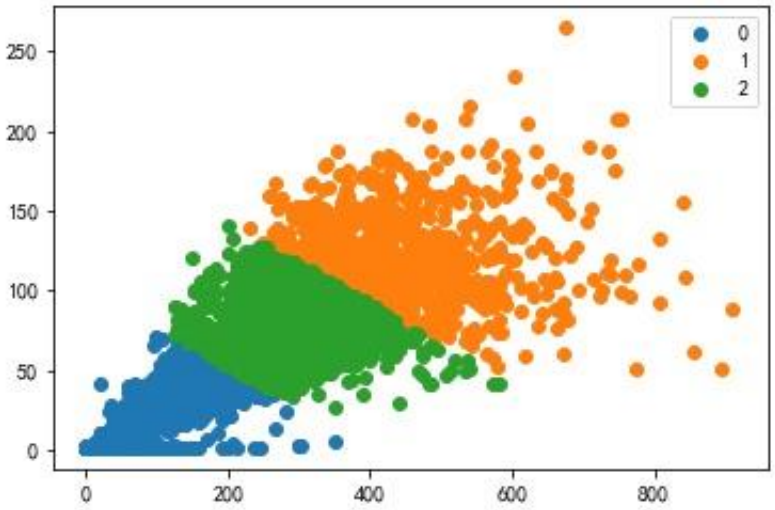


可以看出大部分专业男生的消费总额都略大于女生的消费总额。

3.2 任务 3.2 的解决

任务分析：在此次事例中，需要衡量学生的消费特点，可借用 RFM 模型。RFM 模型是衡量客户价值和客户创利能力的重要工具和手段。与本次任务要求相近，故可用消费频次和消费总额来评判学生的经济情况。

任务方法：对 CardNo 进行分组，计算每个人的消费频次和消费总额，对其进行均值方差标准化，并用这两个指标作为聚类标准，给每个学生贴上对应的标签，并根据标签绘制聚类图。



计算可得聚类中心分别为[120.56，27.51]，[287.71，75.02]，[457,51，121.29]

在[120.56，27.51]聚类中心附近的是 label=0 的同学，此类同学经济较差，学校可适当对这部分同学给予部分补贴，在[287.71，75.02]聚类中心附近的是 label=2 的同学，此类同学经济良好，在[457,51，121.29]聚类中心附近的是 label=1 的同学，此类同学经济富裕。（注：标签并没有顺序之分，在实验中有一个点过于离群，影响聚类效果，已剔除。）