任务的分析和方法

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。

2 任务二：数据分析与可视化

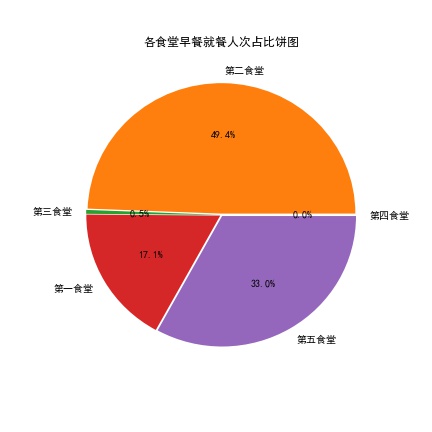
2.1 任务2.1的解决

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。

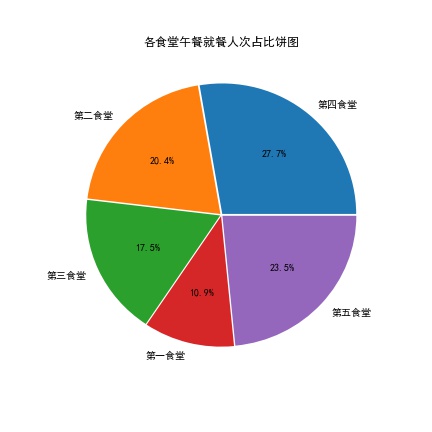
各食堂早餐的就餐人次的饼图如下：



可看出就餐人次占比第二食堂占比最大，其次是第五食堂，然后是第一食堂，第三和第四食堂早上基本没有人去就餐，很有可能是因为第三和第四食堂不提供早餐供应。

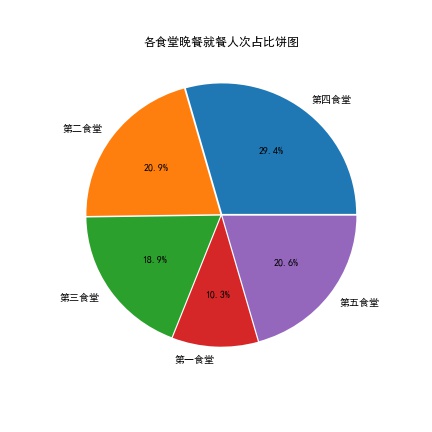
各食堂午餐、晚餐就餐人次绘制方法同早餐就餐人次。

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：



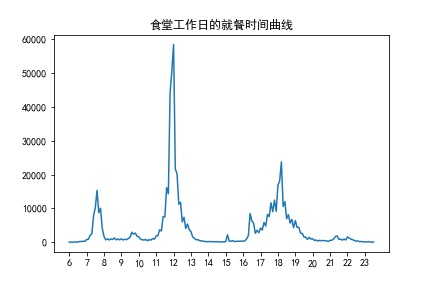
可以看出占比大小顺序为第四食堂>第二食堂>第五食堂>第三食堂>第一食堂。午餐和晚餐食堂就餐人次比例相近，其中第四食堂午餐，晚餐人数最多，深受学生喜爱，不提供早餐。第二，第五食堂一整天都有较大的人流量。第一食堂整天都只有比例较小的人流量，学生最不喜欢的食堂，第三食堂不提供早餐的供应，人流量仅高于第一食堂。

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。

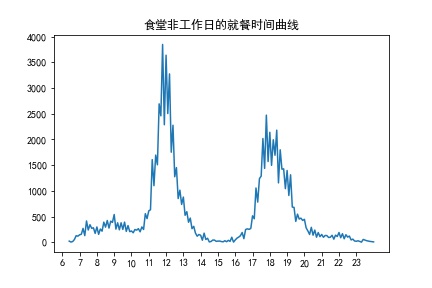
任务做法：用groupby对分钟分组，绘制每分钟的Money总数绘制折线图（消费越高，就餐人数就会越多）。

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：

1. ：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。
2. ：可以在食堂三和食堂四增设早餐的服务。
3. ：在11点前准备好午餐，在17点前准备好晚餐。
4. ：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

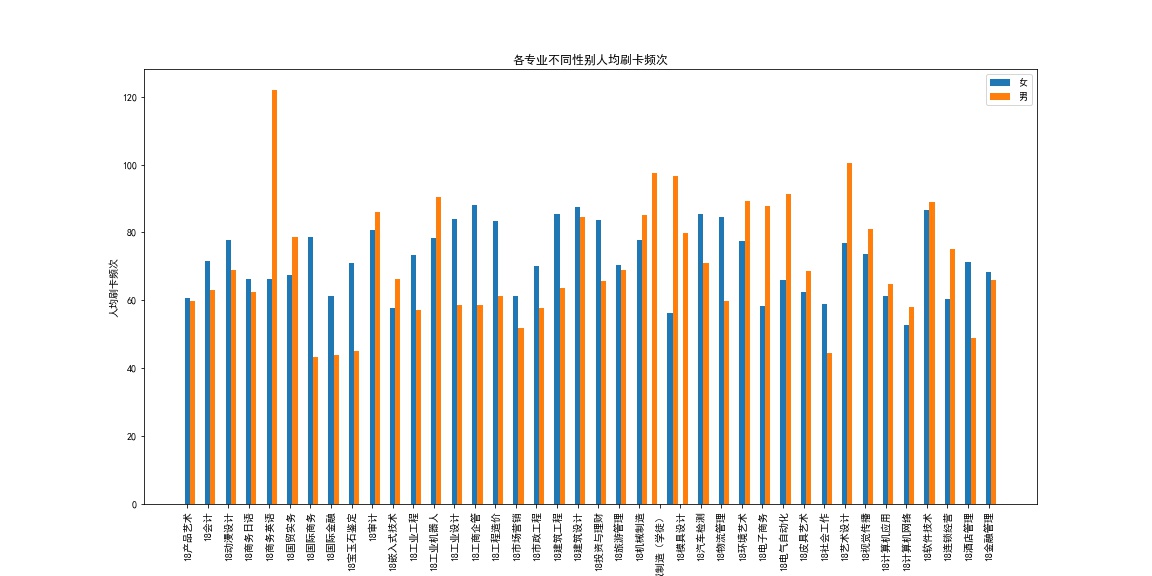
3.1 任务3.1的解决

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。

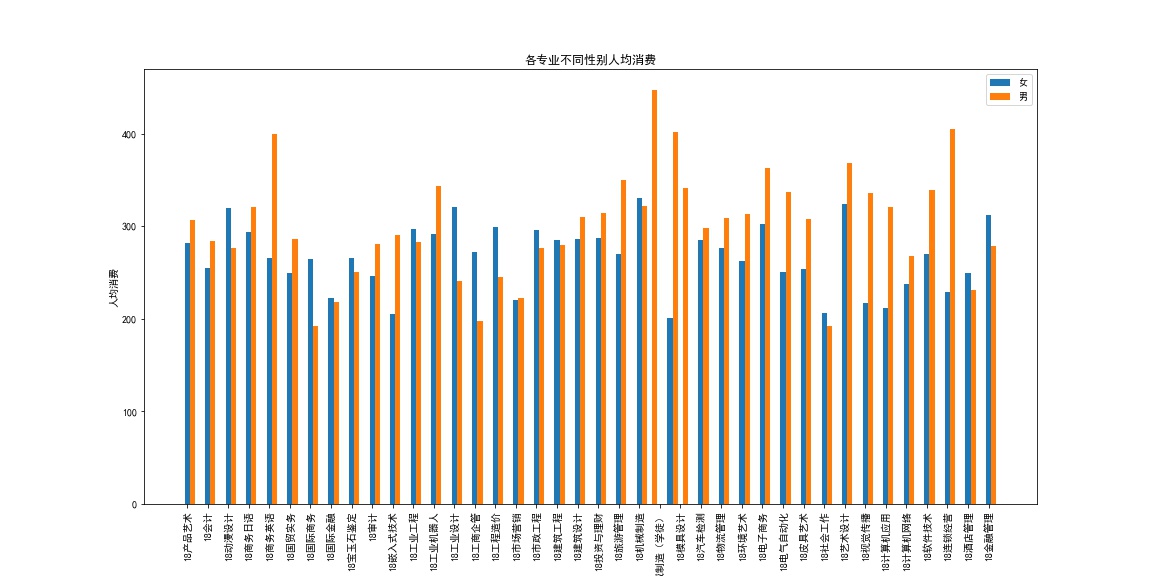
任务②做法：对Major和Sex进行分组，得到每个专业不同性别的消费总额和消费频次，再计算每个专业不同性别的人数，就可以计算不同专业不同性别学生群体的人均刷卡频次和人均消费。

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;

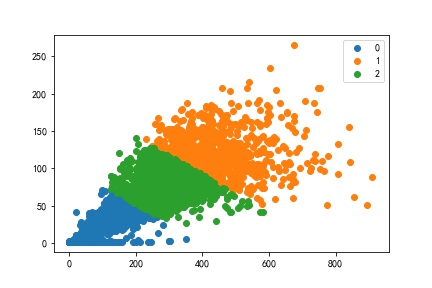


可以看出大部分专业男生的消费总额都略大于女生的消费总额。

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。）