#### 学生校园消费行为分析报告

1. **项目分析**
2. **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。即学生就餐行为分析。

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。即学生消费行为分析。

1. **项目内容**

任务1：数据探索与预处理

* 1. 根据实际项目需求对数据进行必要的数据探索和数据预处理
  2. 对学生个人信息表和消费记录表进行关联

任务2：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 否有显著差别

2.2 分别绘制工作日与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3：学生消费行为分析

3.1 根据18级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，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，为学校判定学生的经济状况提供 参考意见

1. **数据探索和预处理**
2. **任务的完成思路**
3. 任务1.1的完成思路

本任务的内容是根据实际项目的需求对数据进行数据探索与预处理。要想完成任务，首先我们需解决两个问题：对什么数据进行探索？做什么探索与预处理操作？

甲方提供的数据分别为学生ID表与消费记录表，其中消费记录表为主要分析对象，学生ID表可与消费记录表关联，进一步分析学生的具体消费行为。因此我们首先应该单独对两个数据表进行数据探索与数据预处理，在这个过程中，我们需要结合任务2与任务3的内容，筛选我们需要的特征与记录，对数据中的缺失值，重复记录与异常记录做处理，最终将处理后的数据供后续使用。

1. 任务1.2的完成思路

本任务的内容是对学生个人信息表和消费记录表进行关联。本任务的关键点在于选取什么数据进行关联？选用什么方式进行关联？

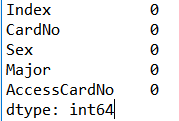
本任务我们应该选取经任务1.1处理后的数据进行关联，选取任务1.1处理后的数据可以使我们不需要再去对合并后的数据进行预处理，减少工作量。选定关联的数据后，观察两个数据表，可以发现数据表中拥有相同的字段(CardNo)，因此我们可以考虑选择基于CardNo进行主键合并。由于数据特性，一个学生可以拥有多个消费记录，但是一个消费记录无法同时属于多个学生，因此我们以学生ID表为左表，消费记录表为右表进行内连接。

1. **任务步骤与过程**
2. 任务1.1
3. 学生ID表的预处理
4. 读取数据

读取指定位置的学生ID表。

1. 处理缺失值

对数据进行缺失值统计，统计结果如下:



由结果表示学生ID表中各个特征的缺失值记录为0，数据表无缺失值，因此无需进行缺失值处理。

1. 处理重复值

在学生ID表中，每个学生的校园卡号都是唯一，因此我们以校园卡号列进行去重操作，表示同一个校园卡号只有唯一的一条信息。

执行结果如下：



由结果可知，去重前后的记录的行列并无变化，可知学生校园卡号列并无重复值，因此可推测学生ID表无重复值。

1. 处理无用特征

在表用存在序号Index，门禁卡号AccessCardNo与校园卡号CardNo三个唯一性的数据，三者冲突，并且Index在表中无实际用处，但是所以我们删除Index特征。AccessCardNo则可以关联相关的门禁信息，但是我们并不分析门禁信息，因此也选择删除AccessCardNo特征。

1. 保存数据

将经过前面步骤处理过后的数据命名为task1\_X1.csv保存到指定位置。

1. 消费记录表
2. 读取数据

读取指定位置的消费记录表。

1. 分析提取有用记录

由于在消费记录表中，存在多种类型的消费数据（存款，发卡存款，取款，退款，无卡销户，消费），其中消费类型的数据是我们分析的主要对象，因此我们提取消费类型的数据进行分析。

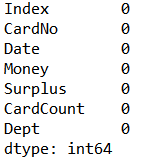
1. 处理无用特征

由于我们任务是对于食堂就餐行为与学生消费行为进行分析，因此消费记录表中消费项目的序列号TermSerNo与消费操作的编码conOperNo等可能用于关联其他数据表格的特征，我们将其剔除，以免影响代码执行效率。而在其他字段特征中，校园卡号CardNo与校园卡编号PeoNo都能验证学生的唯一性，且数据重复，但由于我们后续可能需要校园卡号CardNo进行多表连接，因此我们选择剔除校园卡遍号PeoNo特征。此外，由于我们在上面筛选了消费类型Type的数据进行分析，因此Type特征的值都为消费而存储金额FundMoney的值都为0，所以我们顺带把Type特征和FundMoney特征也剔除。

因此我们总共需要去除四个特征：校园卡遍号PeoNo，消费项目的序列号TermSerNo，消费操作的编码conOperNo，消费类型Type和存储金额FundMoney。

1. 处理缺失值

对数据进行缺失值统计，统计结果如下:



由结果表示消费记录表中各个特征的缺失值记录为0，数据表无缺失值，因此无需进行缺失值处理。

1. 处理重复值

对于重复值的统计，由于订单号Index是消费记录的唯一标识，因此我们只需要判断订单号Index是否重复，即可判断所有记录是否有重复值。

执行结果如下：

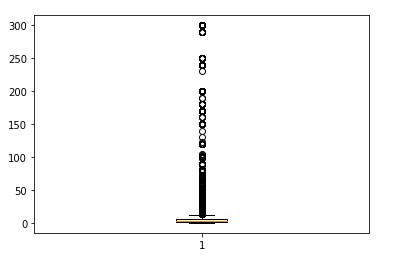


由结果可知，去重前后的记录行列没有发生变化，因此我们可以推断，数据中的所有记录不存在重复。

1. 异常值分析

由于我们的任务主要针对学生的消费行为进行分析，而在学生的消费行为中，消费金额Money这个特征是最为主要的数值型特征，因此我们先对Money特征画箱线图分析数据是否存在异常。

箱线图如下：



由箱线图可知，数据中并无小于0的消费金额，而图中部分数据高于上限，这部分数据可能是异常值，也可能是部分高消费同学的记录，因此我们无法具体判断是否为异常值去除。

其次，我们对日期消费时间进行分析，正常的校园消费应该是发生在早上6：00-晚上23：00期间的，基本不存在会有半夜进行的消费，因此我们需要判断是否存在半夜消费并去除。

1. 保存数据

将经过前面步骤处理过后的数据命名为task1\_X2.csv保存到指定位置。

1. 任务1.2
2. 读取数据

读取指定位置中经过任务1.1处理后的学生ID表task1\_X1.csv与消费记录表task1\_X2.csv。

1. 主键合并

在学生个人信息表中存在一个唯一性的特征校园卡号CardNo，而在消费记录表中同样存在CardNo这个标识学生的特征，因此我们对两个表以CardNo进行主键关联。

1. 处理无用特征

在合并数据表后，消费记录表中的Index特征已经失去的原先的作用，因此我们将此特征去除。

1. 保存数据

将合并后的数据命名为task1\_1\_X.csv保存到指定位置。

1. **食堂就餐行为分析**
2. **任务的完成思路**
3. 任务2.1的完成思路

本任务的内容是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显著差别。这个任务的关键点在于如何计算就餐人次？早中晚餐时间区域的划分？

首先，我们需要理解就餐人次的概念，即在同一地点，同一时间范围内的多次刷卡记录看作一次就餐行为，也就是一次就餐人次。

其次，我们需要对早中晚餐的时间区域进行划分。此处，我们按照校园学生的生活习惯将7点-9点的时间段划分为早餐时间，11点-13点的时间段划分为午餐时间，18点-19点的时间划分为晚餐时间。

最后，我们只需提取对各食堂早中晚餐时间区域的就餐人次进行频数统计，绘制早中晚餐就餐区域占比饼图即可。

1. 任务2.2的完成思路

本任务的内容是分别绘制工作日与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。这个任务的关键点在于如何划分工作日与非工作日？

由于甲方提供给我们的数据是2019年4月份的消费记录，因此我们只需要去查找2019年4月中的工作日与非工作日，根据日期进行划分。

划分完工作日与非工作日后，我们分别计算工作日与非工作日的记录每小时的就餐频次，绘制出对应的就餐时间曲线图。

1. 任务2.3的完成思路

本任务的内容是根据上述分析的结果，为食堂的运营提供建议。因此我们需要结合上述两个任务的结果，综合分析学生的就餐规律，为食堂提供建议。

1. **任务步骤与过程**
2. 任务2.1的步骤与过程
   1. 读取数据

读取指定位置中经过任务1.1处理后的消费记录表task1\_X2.csv。

* 1. 数据排序

对读取的数据按校园卡号CardNo与消费时间Date进行升序排序，方便后面就餐人次的标记与计算。

* 1. 提取时间序列

对数据进行操作，将消费时间Date列转化为DateTime类型，进一步提取时间序列Week、Day、Hour与Minute。

* 1. 筛选食堂记录

对经过前面步骤的数据进行筛选，筛选消费地点包含食堂的记录，消费地点中包含食堂的地方分别为第一食堂、第二食堂、第三食堂、第四食堂、第五食堂与教师食堂共6个地点。将筛选出来的地方分类保存到列表中。

* 1. 标记就餐人次

首先我们为消费记录新增一列lab，并设置为0。其次遍历食堂的记录，在循环中将前后行为不同校园卡号与不同天的记录进行标记，将后一行的lab特征设置为1。然后进行一步筛选，将前后行记录中消费时间大于10分钟的记录进行标记，将后一行的lab特征设置为1。

* 1. 划分早中晚餐的消费记录集

遍历不同食堂的消费记录，根据时间序列Hour进行划分，将Hour>7且Hour<9的记录划分为早餐消费记录，将Hour>11且Hour<13的记录划分为午餐消费记录，将Hour>18且Hour<20的记录划分为晚餐消费记录，将记录保存起来。

* 1. 计算各食堂的累计就餐次数

遍历不同食堂不同时段的消费记录，统计消费记录中lab值为1的次数，该次数即为该时段食堂累计就餐次数，将累计的就餐次数保存至列表sums。

* 1. 绘制就餐人次饼图

根据上述统计的不同时段不同食堂的就餐次数进行绘图，将绘制结果保存至本地指定位置。

1. 任务2.2的步骤与过程
   1. 读取数据

读取指定位置中经过任务2.1处理后的消费记录表task2\_X1.csv。

* 1. 提取工作日与计算工作日天数

由于在4月中，原本为非工作日的28号因为调休变成了工作日而原本为工作日的5号变成了非工作日，此外所有周一至周五的记录都是工作日。因此我们首先提取数据中Day等于28的记录，在提取Day不等于5并且Week小于等于5的记录，提取的两个类记录进行合并。最后计算合并后的数据中Day列不重复的元素的个数，即为工作日天数。

* 1. 计算工作日所有食堂每小时的就餐频数

累计数据中不同Hour值中lab特征为1的个数，将结果保存到workday\_sum列表中。

* 1. 划分食堂并计算工作日各食堂每小时的就餐频数

对经过前面步骤的数据进行筛选，筛选消费地点包含食堂的记录，消费地点中包含食堂的地方分别为第一食堂、第二食堂、第三食堂、第四食堂、第五食堂与教师食堂共6个地点。将筛选出来的地方分类保存到列表workday\_data3中。

遍历workday\_data3，累计同一食堂不同Hour值中lab特征为1的个数，将结果保存到hour\_sum列表中，最后将六个食堂的hour\_sum列表添加到hours\_sum列表。

* 1. 绘制食堂工作日日均就餐人次曲线图

绘制食堂工作日日均就餐人次曲线图，首先计算各个食堂每小时的日均就餐人次（即每小时的工作日就餐人次/工作日天数），其次，计算工作日所有食堂每小时的的日均就餐人次的平均值（即每小时的工作日就餐人次/工作日天数/6），最后，以后小时为x周，以各个食堂与平均值为y轴绘制食堂工作日日均就餐人次曲线图。

* 1. 提取非工作日与计算非工作日天数

由于在4月中，原本为工作日的5号因为调休变成了工作日而原本为非工作日的28号变成了工作日，此外所有周末的记录都是非工作日。因此我们首先提取数据中Day等于5的记录，在提取Day不等于28并且Week大于5的记录，提取的两个类记录进行合并。最后计算合并后的数据中Day列不重复的元素的个数，即为非工作日天数。

* 1. 计算非工作日所有食堂每小时的就餐频数

累计数据中不同Hour值中lab特征为1的个数，将结果保存到unworkday\_sum列表中。

* 1. 筛选食堂记录并计算非工作日各食堂每小时的就餐频数

对经过前面步骤的数据进行筛选，筛选消费地点包含食堂的记录，消费地点中包含食堂的地方分别为第一食堂、第二食堂、第三食堂、第四食堂、第五食堂与教师食堂共6个地点。将筛选出来的地方分类保存到列表unworkday\_data3中。

遍历unworkday\_data3，累计同一食堂不同Hour值中lab特征为1的个数，将结果保存到hour\_sum列表中，最后将六个食堂的hour\_sum列表添加到hours\_sum列表。

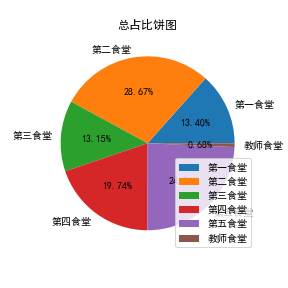
* 1. 绘制食堂非工作日均就餐人次曲线图

绘制食堂非工作日日均就餐人次曲线图，首先计算各个食堂每小时的日均就餐人次（即每小时的非工作日就餐人次/非工作日天数），其次，计算非工作日所有食堂每小时的的日均就餐人次的平均值（即每小时的非工作日就餐人次/非工作日天数/6），最后，以后小时为x周，以各个食堂与平均值为y轴绘制食堂非工作日日均就餐人次曲线图。

1. **任务结果及分析**
2. 任务2.1结果及分析

任务2.1中总绘制出4张食堂占比饼图。

第一张为总占比饼图，即不同食堂一整天学生消费占比饼图。饼图如下：

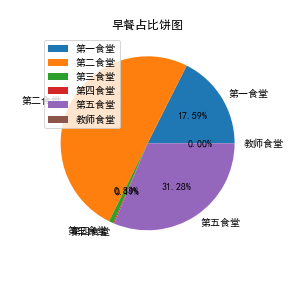


分析：

从饼图中我们可以看出，第二食堂的占比最大，其次是第五食堂，接下来是第四食堂，第一食堂，第三食堂，最后是教师食堂。

由此可以看出，学生们倾向往第二食堂与第五食堂消费，而教师食堂却很少有学生过去消费。

第二张为早餐占比饼图，饼图如下：

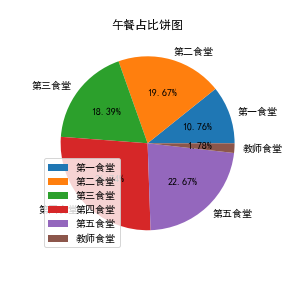


分析：

从饼图中我们可以看出，第二食堂的占比最大，其次是第五食堂，接下来是第一食堂，第三食堂，第四食堂，最后是教师食堂。

由此可以看出，学生早餐的就餐地点基本分布在第二食堂，第五食堂与第一食堂，而第三第四食堂则只有少部分学生会进行消费，教师食堂基本没有学生进行消费。

第三张为午餐占比饼图，饼图如下：

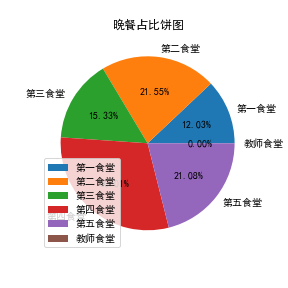


分析：

从饼图中我们可以看出，第四食堂的占比最大，其次是第五食堂，接下来是第二食堂，第三食堂，第一食堂，最后是教师食堂。

由此可以看出，大部分学生会选择午餐在第四食堂或第五食堂消费，较少学生会选择第一食堂，极少学生会选择教师食堂。

第四张为晚餐占比饼图，饼图如下：



分析：

从饼图中我们可以看出，第四食堂的占比最大，其次是第二食堂，接下来是第五食堂，第三食堂，第一食堂，最后是教师食堂。

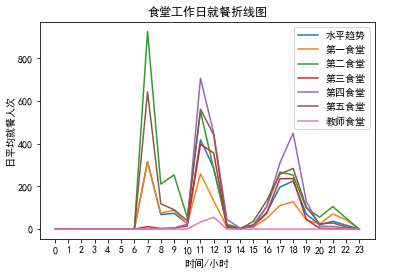
由此可以看出，学生晚餐的就餐地点基本分布在第四食堂，第五食堂与第二食堂，而第一第三食堂则只有少部分学生会进行消费，教师食堂基本没有学生进行消费。

综上所述，学生早餐的就餐地点与午餐晚餐存在较大的差异，而午餐与晚餐差别不大。

1. 任务2.2结果及分析

任务2.2总共绘制出两张曲线图。

第一张为食堂工作日就餐趋势曲线图，曲线图如下：



就餐峰值分析：

第一食堂：在7点左右迎来了早餐就餐峰值，就餐人次约为300，在11点左右迎来午餐就餐峰值，就餐人次约为250，在18点左右则迎来了晚餐的就餐峰值，就餐人次约为120。

第二食堂：在7点左右迎来了早餐就餐峰值，就餐人次约为900，在11点左右迎来午餐就餐峰值，就餐人次约为550，在17点左右则迎来了晚餐的就餐峰值，就餐人次约为250。

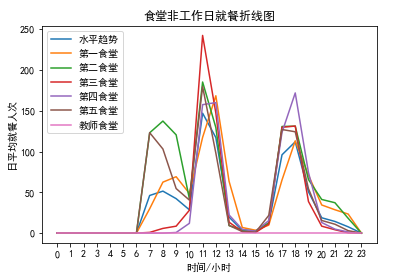
第三食堂：在7点左右迎来了早餐就餐峰值，就餐人次约为10，在11点左右迎来午餐就餐峰值，就餐人次约为400，在17-18点左右则迎来了晚餐的就餐峰值，就餐人次约为240。

第四食堂：第四食堂工作日早餐时间就餐人次基本为0，在11点左右迎来午餐就餐峰值，就餐人次约为690，在18点左右则迎来了晚餐的就餐峰值，就餐人次约为420。

第五食堂：在7点左右迎来了早餐就餐峰值，就餐人次约为640，在11点左右迎来午餐就餐峰值，就餐人次约为550，在18点左右则迎来了晚餐的就餐峰值，就餐人次约为260。

教师食堂：食堂工作日早餐时间就餐人次基本为0,在12点左右迎来午餐就餐峰值，就餐人次约为40，食堂工作日晚餐时间就餐人次基本为0。

第二张为食堂非工作日就餐趋势曲线图，曲线图如下：



就餐峰值分析：

第一食堂：在9点左右迎来了早餐就餐峰值，就餐人次约为65，在12点左右迎来午餐就餐峰值，就餐人次约为160，在18点左右则迎来了晚餐的就餐峰值，就餐人次约为120。

第二食堂：在8点左右迎来了早餐就餐峰值，就餐人次约为140，在11点左右迎来午餐就餐峰值，就餐人次约为250，在17点左右则迎来了晚餐的就餐峰值，就餐人次约为125。

第三食堂：第三食堂非工作日早餐时间无明显就餐峰值，在11点左右迎来午餐就餐峰值，就餐人次约为240，在17-18点左右则迎来了晚餐的就餐峰值，就餐人次约为135。

第四食堂：第四食堂非工作日早餐时间无明显就餐峰值，在11-12点左右迎来午餐就餐峰值，就餐人次约为165，在18点左右则迎来了晚餐的就餐峰值，就餐人次约为170。

第五食堂：在7点左右迎来了早餐就餐峰值，就餐人次约为120，在11点左右迎来午餐就餐峰值，就餐人次约为180，在17-18点左右则迎来了晚餐的就餐峰值，就餐人次约为130。

教师食堂：教师食堂非工作日全天时间就餐人次基本为0。

1. 任务2.3结果及分析

综合以上图论，我们给食堂以下一些建议：

第一食堂：在非工作日中，第一食堂的销售情况远比工作日要好，建议该食堂增加非工作日的营业。

第二食堂：大部分学生早餐会选择第二食堂，建议该食堂可以适量增加早餐的供应量。第二食堂的工作日就餐情况也非工作日要好。

第三食堂：由于只有极少部分同学早餐会选择在第三食堂进行，建议该食堂可以减少早餐的供应量。

第四食堂：由于只有极少部分同学早餐会选择在第四食堂进行，建议该食堂可以减少早餐的供应量，而大部分同学午餐晚餐选择在第四食堂，建议该食堂可以适量增加早餐的供应量。

第五食堂：在非工作日中，第五食堂的销售水平高与平均水平，而正在工作日则与平均水平相差不大。

教师食堂：由于教师食堂的主要性质是供教师进餐而非供学生，因此无法做建议。

1. **学生消费行为分析**
2. **任务的完成思路**
   1. 任务3.1的完成思路

任务3.1的内容是根据18级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，分析不同专业间不同性别学生群体的消费特点。由于我们分析的对象是18级整体学生，因此我们需选择在任务1.2关联后的数据表进行操作。其次是刷卡频次与人均消费的计算，需要我们计算去每个学生的刷卡频次与人均消费。最后我们可以通过柱形图去展示各个专业中男女生的消费特点。

（2） 任务3.2的完成思路

任务3.2的内容是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，为学校判定学生的经济状况提供参考意见。任务的关键在于选择什么特征作为指标构建去构建聚类？

我们参考消费行为模型RFM模型进行指标的选取，RFM模型中有三个指标分别为R（最近一次消费）、F（消费频率）、M（消费金额）。而在我们实际项目中，分析的学生最近一次消费并无法达到我们的要求，反而会影响我们的聚类结果，因此我们选择FM作为我们的聚类指标，即学生的刷卡频率与消费金额。我们使用K-Means算法去构建模型，构建完模型后我们ARI评价法（兰德系数）对模型进行评价。选出最优的聚类组数，绘出聚类效果图。

1. **任务步骤与过程**
   1. 任务3.1的步骤与过程
      1. 读取数据

读取指定位置中经过任务1.2处理后的消费记录表task1\_1\_X1.csv。

* + 1. 计算18级整体学生本月的人均刷卡频次与人均消费

由于我们获取的时候18级整体学生关联消费记录的数据，因此，我们只需要使用groupby对CardNo进行分组，并计算每一组Money的总数，即可获得18级学生人总消费金额。再对18级学生人总消费金额求均值，即可获得18级整体学生本月的人均消费。人均刷卡频次同理，对CardNo做分组，计算每一组记录的总数。再对所有组的总数记录求均值，即可获得18级整体学生本月的人均刷卡频次。

* + 1. 获取专业列表

我们对步骤a获取的数据以Major做去重操作，并提取处理后的数据的Major列的值，保存到列表majors。

* + 1. 划分不同专业不同性别的学生数据

首先我们需要对不同专业的学生数据做筛选分组，将结果保存到列表majors\_data。其次，我们对数据做进一步的划分，同一专业的数据按性别划分，将结果保存到列表major\_sex\_data中，最后将不同专业的major\_sex\_data列表保存至majors\_sex\_data列表中。

* + 1. 计算不同专业不同性别学生人均刷卡频次与人均消费

首先遍历每个专业中不同性别学生的数据，分别对数据进行一次步骤b的操作，分别将结果保存至consume\_sex\_data与card\_sex\_data，最后将所有专业的sonsume\_sex\_data与card\_sex\_data数据保存到consumes\_sex\_data与cards\_sex\_data列表。

* + 1. 绘制柱形图

以每个专业的consume\_sex\_data列表为y值，以专业的学生性别为x轴绘制柱形图，保存至本地。

* 1. 任务3.2的步骤与过程
     1. 读取数据

由于我们分析的学生整体的校园消费行为，因此我们读取整体学生的消费记录。读取指定位置中经过任务1.1处理后的消费记录表task1\_X2.csv。

* + 1. 统计每个学生的月消费频率与消费金额，生成新数据表

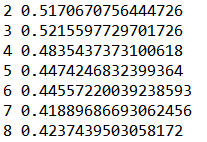
使用groupby对数据按CardNo做分组，分别计算每一组Money的总和与每一组CardCount的总数。对分组后的两个GroupBy对象重新设置索引，并合并为一个数据表。

* + 1. 标准差标准化数据

提取新数据表中的CardCount列与Money列，对其进行标准差标准化处理，结果保存在model\_data\_ss列表中。

* + 1. ARI评价法（兰德系数）选择最优组

构建一个2->8的循环，测试不同组数构建的KMeans聚类模型的兰德系数。结果如下：



由于兰德系数是数值越靠近1越好，因此我们将聚类组数设置为3

* + 1. 绘制最优组的聚类效果图

以3为组数重新进行KMeans聚类，以聚类后的模型去绘制散点图，将结果保存至本地。

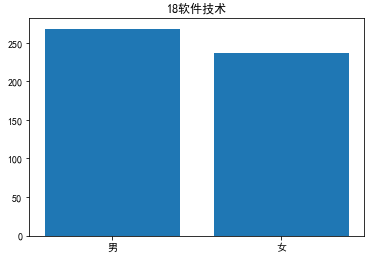
1. **任务结果及分析**
   1. 任务3.1的结果及分析
   2. 18级整体学生本月人均刷卡频次和人均消费

结果如下：



* 1. 各专业男女生消费对比柱形图（以18软件技术的学生为例）

结果如下：



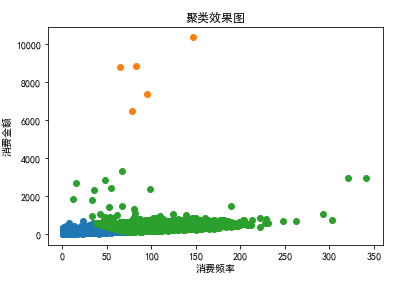
分析：

由柱形图可知，18级软件技术的男生消费水平高于18级软件技术的女生。男生的消费水平约为285，女生的消费水平约为235。

相较18级整体学生的人均消费，男生的消费水平在整体学生的消费水平左右。而女生则低于整体学生的消费水平。

* 1. 任务3.2的结果及分析

任务3.2绘制了一张聚类效果散点图，散点图如下：



分析：

聚类的模型共分为三类，按消费频率与消费金额的比例分为低消费，中消费与高消费。当消费频率与消费金额较低时，为低消费；当消费频率高，消费金额低或者消费金额高消费频率高时，为中消费；当消费频率低与消费金额高时，为高消费。

由散点图可以看出，蓝色区域为低消费者，绿色区域为中消费者，橙色区域为高消费者。

因此，当学生的状况满足对应区域时，就可判断学生处于哪个消费水平。