学生校园消费数据分析

[一、项目背景 2](#_Toc28183808)

[二、分析思路 2](#_Toc28183809)

[三、操作过程 3](#_Toc28183810)

[（一）、数据探索及预处理 3](#_Toc28183811)

[1、了解获得的数据集大小及变量类型 3](#_Toc28183812)

[2、对重复值和空值进行探索和处理 3](#_Toc28183813)

[3、删除多余的列： 4](#_Toc28183814)

[4、提取消费型数据： 4](#_Toc28183815)

[5、查看异常值并进行处理： 4](#_Toc28183816)

[6、对消费数据进行整合： 5](#_Toc28183817)

[7、根据任务目标获取不同的数据集并保存： 6](#_Toc28183818)

[（二）、对食堂数据进行分析： 7](#_Toc28183819)

[1、早中晚各食堂就餐人次占比饼图绘制 7](#_Toc28183820)

[2、工作日与非工作日就餐时间曲线绘制 7](#_Toc28183821)

[（三）、对18级消费数据进行分析 7](#_Toc28183822)

[1、对整体数据进行分析 7](#_Toc28183823)

[2、对不同专业不同性别族群进行分析 7](#_Toc28183824)

[3、对不同消费类型族群进行分析 8](#_Toc28183825)

[四、分析结果 11](#_Toc28183826)

[（一）、食堂数据分析及建议 11](#_Toc28183827)

[1、早中晚各食堂就餐人次占比饼图结果分析： 11](#_Toc28183828)

[2、就占比饼图给学校食堂运营的建议： 13](#_Toc28183829)

[3、工作日与非工作日就餐曲线图： 14](#_Toc28183830)

[4、以就餐时间曲线给学校食堂运营的建议： 15](#_Toc28183831)

[（二）、18级学生消费数据分析及建议 15](#_Toc28183832)

[1、 对整体消费行为进行分析 15](#_Toc28183833)

[2、对不同专业不同性别族群进行分析 15](#_Toc28183834)

[3、对不同消费类型族群进行分析 18](#_Toc28183835)

[4、给学校判定学生的经济状况的参考意见： 20](#_Toc28183836)

# 一、项目背景

校园一卡通作为集身份认证、金融消费、数据共享等多项功能于一体的信息集成系统，在为师生提供服务的同时，系统自身也累积了大量的历史纪录，其中蕴含着学生的消费记录以及学习食堂等各部门的运行状况信息。因此，对校园一卡通数据进行分析，有利于了解学校服务及相关部门的运行情况及为改进学校服务并为相关部门提供决策信息提供支持。

在本份报告中，将根据某高校4月份的校园一卡通数据及18级学生个人信息数据，分别对学校食堂运营情况及18级学生消费行为，利用数据分析及建模的方法，进行信息挖掘，分析学生在校园内的学习生活行为，并为改进学校服务及相关部门运营提供建议。

# 二、分析思路

从分析内容可知，我们需要针对学校食堂及18级学生消费情况进行分析，因此在数据分析中我们需要分别提取**学校食堂消费数据**和**18级学生消费数据**。

针对**食堂消费数据**，本次数据分析中将通过绘制早中晚各食堂就餐人次占比饼图，以及工作日与非工作日就餐时间曲线进行分析。因此，在分析中需要确定早中晚餐的就餐时间段和工作日与非工作日的划分；此外，将同一时间间隔、同一地点、同一张卡的消费数据进行整合。最后对数据进行整理解读，结合分析结果给食堂运营提供建议。

针对**18级学生消费数据**，本次分析中将从总体分析，不同专业不同性别学生消费行为比较，及利用聚类对不同消费特征群体进行比较，三个方面进行分析。因此，需要先计算月人均刷卡频次和人均消费 。接着，利用专业和性别将这些消费数据进行分组，统计不同专业不同性别学生群体的月均消费金额和频数，进行比较分析 。然后，基于RFM模型，利用KMeans聚类算法选，取月总消费频次、月均笔消费和周末均笔消费对学生消费行为进行分类 ，并描述不同族群学生的消费特点。最后，根据分析的结果，给学校判定学生的经济情况提供参考建议 。

# 三、操作过程

## （一）、数据探索及预处理

在该部分中，需要对将进行分析的数据进行数据探索及预处理，以了解这些数据的特征，并将其中的无用值、异常值进行清理，最后根据任务目标对数集进行重新提取和划分。

### 1、了解获得的数据集大小及变量类型

经查看，文件：data2.csv –某高校4月一卡通数据集，为11列，519367行的数据集。其变量为：'CardNo', 'PeoNo', 'Date', 'Money', 'FundMoney', 'Surplus', 'CardCount','Type', 'TermSerNo', 'conOperNo','Dept'

文件：data1.csv—某高校18级学生信息表， 为4列，4341行的数据集。其变量为：'CardNo', 'Sex', 'Major', 'AccessCardNo'

由此可见，**两份表格都具有'CardNo' 变量**，可作为之后提取18级学生消费数据集的主键。从文件data2.csv 的行数可见，该数据集中**含有519366条消费记录**；从文件data1.csv可见该高校**在册18级学生共有4330位**。

### 2、对重复值和空值进行探索和处理

2.1、清理完全重复的记录：

利用drop\_duplicates( ) 函数对文件data2的完全重复数据进行清理。最终将其中3条完全重复消费记录清除。

2.2、清理卡号和刷卡频次完全重复的记录：

经检验，文件data2的数据集中，一共有 205对 记录是重复的。将这些记录提取出来发现，每对记录可以看作是卡号相同的两张卡在进行独立消费。而在整个分析中，我们假设一个卡号只代表一张卡，且该卡的消费记录只代表一个人的消费记录；此外，考虑到这些数据的占比不足1%，因此利用drop\_duplicates(subset=[‘CardNo’,’CardCount’])将每对记录的其中一条删去。此时，数据集中有519159条记录。

2.3、对空值进行探索：

利用print(info.isnull().sum())，发现列TermSerNo 和列 conOperNo仅有较少样本具有数值，因此先对这两列数据进行分析再进行处理。

2.3.1、**对TermSerNo 列数据进行探索**：发现该列非空的数据个数为 7259 个，占经初步处理后整体数据的1.40%。先对其进行抽样，选取其中20个数据，经观察发现这些数据都有的一个特点：在**凌晨0：00：00进行消费**。接着，对这些非空数据进行探索，发现所有数据都是凌晨0：00：00的消费记录。而凌晨0：00：00的消费数据将在之后异常值处理中进行处理，由此判断TermSerNo的值暂无其他价值，将这一整列数据删去。

2.3.2、**对conOperNo列数据进行探索**：发现该列非空的数据个数为251个，占经初步处理后整体数据的0.05%。同样的先对其进行抽样，选取其中20个数据，经观察发现抽样数据都有的一个特点：**不为“消费”型数据**。接着，对这些非空数据进行探索，发现所有数据都不为“消费”型数据。而我们由于需要进行数据分析的数据应为消费型数据，所以接下来将会进行“消费”型数据的提取。因此，此时TermSerNo的值暂无其他价值，将这一整列数据删去。

### 3、删除多余的列：

经检验，变量 PeoNo ，FundMoney和Type 对接下来的分析没有作用，为了达到更好的运行效率，计划删去这三列的所有数据。

### 4、提取消费型数据：

由于整个数据分析是针对消费数据进行分析，因此只提取类型为消费型的数据。此时，data2 **数据集有500548条数据**。

### 5、查看异常值并进行处理：

5.1、从消费时间角度对数据进行预处理，排除异常值：

为了方便以消费时间为依据进行数据处理，从数据集提供的 ’Date’ 变量中，**提取日期和时间**做新的两列进行之后的数据分析，并按 ’Date’ 的时间先后由最开始记录的数据依次进行排列，之后删去 ’Date’ 列。

在本报告中，假设的学校部门的**营业时间为6：00-24：00**（时间型数据精确到分钟）。因此，0：00-5：59的消费记录为非营业时间的消费记录，可判断为异常值；此外，由于前面对TermSerNo值非空的数据的判断，我们在处理前，需要对0：00消费数据进行探索。

经检验，零点消费数据共有7259个，和TermSerNo值非空的数据个数重合，由此判断零点消费数据都为TermSerNo值非空的数据，可以判定这些消费数据都是异常数据，在之后的处理中会和其他非营业时间的消费记录一起删去。

在该阶段处理中，将0：00-5：59的消费数据删去。此时，**data2数据集有492316条记录**。

5.2、从消费额度异常进行分析，排除异常值：

通过箱线图对消费额度进行分析。获得单笔消费金额箱线图如下显示：

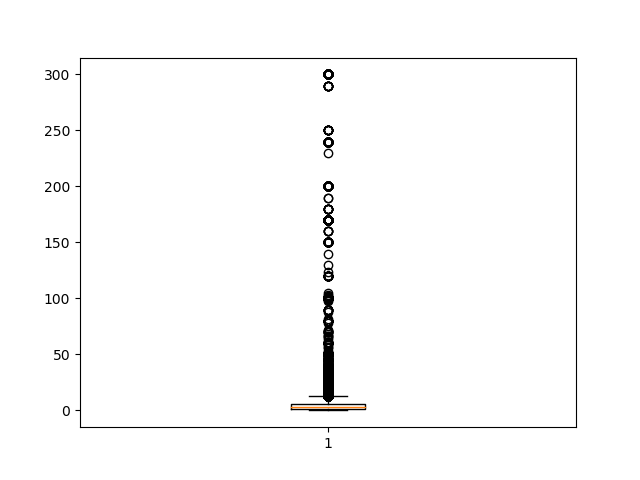


图1 单笔消费金额箱线图（单位：元）

从分析结果可见，虽然**箱线图的值域远低于50**，但是考虑到有单次高消费的情况及值在100以下仍旧有较高密度，因此将考察大于100的值。经抽样考察，这些值都没有异常的情况，因此从消费金额角度考察，数据集中没有异常值，不进行异常值处理。

### 6、对消费数据进行整合：

考虑到同一个人，在同一个学校部门，短时间内进行多次消费的情况，在分析中需要将同一时间段、同一地点和同一张卡的消费进行合并。

首先，对“短时间”这个概念进行量化。计算同一张卡在同一个部门内未进行消费地点转移时所有消费记录的最大时间差，绘制箱线图进行分析。箱线图如图所示：

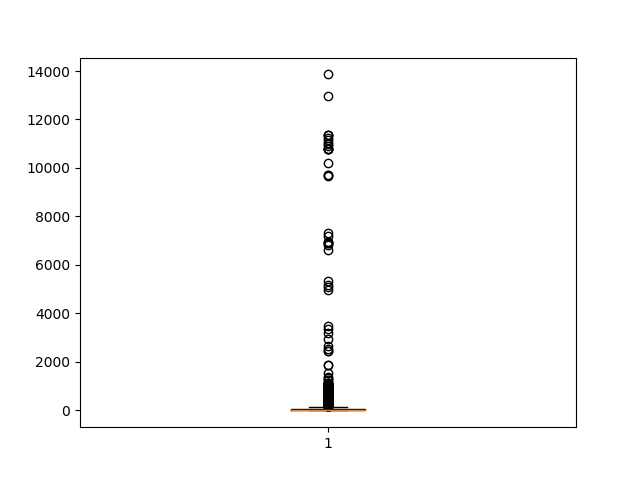


图2 均笔消费金额箱线图（单位：秒）

从图2中可见，在2000s以下为一个密集记录段，结合生活经验，以**30分钟**做该指标的定量标准。

在数据合并的过程中，将**同一张卡，在同一地点（未发生消费地点转移），30分钟内的消费**合并未一次消费，重新计算消费金额。同时，为了之后以时间为x轴绘制就餐曲线图，将所有记录的时间点按时间段再标记。其中为了统计方便且具有代表性，以**20分钟**为一时段进行数据合并。

将整合后的数据集保存为文件：task1\_X2.csv

### 7、根据任务目标获取不同的数据集并保存：

7.1、获取食堂消费数据集并保存：

选取task1\_X2.csv 数据集中，**变量‘Dept’为：'第一食堂','第二食堂','第三食堂','第四食堂','第五食堂',’教室食堂’**的数据进行数据集划分，从而获得学校食堂消费数据集。将提取的数据集保存为文件：task1\_X3.csv

7.2、获取18级消费数据集并保存：

将数据集data1和task1\_X2，**以'CardNo' 做主键合并，并选择内连接方式合并**，从而获得18级消费数据集。将该数据集保存为文件：task1\_X4.csv

（二）、对食堂数据进行分析：

在该部分的分析中，共从两个角度进行分析：早中晚各食堂就餐人次占比和工作日与非工作日就餐时间曲线。因此要分别从这两个角度进行数据再划分和分析。

### 1、**早中晚各食堂就餐人次占比饼图绘制**

按时间节点将数据分组，接着进行早中晚各食堂就餐人次占比的饼图绘制。

此处，假设早餐运营时段为：06：00 - 10：59

午餐运营时段为：11：00 - 13：59

晚餐运营时段为：17：00 - 19：59

### 2、工作日与非工作日就餐时间曲线绘制

按工作日与非工作日对食堂数据进行划分，然后以每个时段的日均就餐人次总数做y轴，运营时段做x轴，进行折线图绘制，并且将工作日与非工作日的就餐曲线图展现在同一张表中。

（三）、对18级消费数据进行分析

在该部分的分析中，将从三个角度：总体，不同专业不同性别的群体和不同消费类型的群体进行分析。因此在数据划分及分析时也应从这三个方向入手。

### 1、对整体数据进行分析

计算4月人均刷卡频次和人均消费金额。

### 2、对不同专业不同性别族群进行分析

将18级消费数据集按不同专业不同性别分组，统计4月人均刷卡频次和人均消费金额，并绘制柱状图进行比较。

### 3、对不同消费类型族群进行分析

利用聚类分析和RFM模型对18级4月消费学生进行归类，并作比较分析。

3.1、基于RFM模型的指标选取

RFM模型是衡量客户价值和客户创利能力最常用的模型之一。其中R 为最近一次消费的时间，F为一段时间内的消费频次，M为一段时间内的消费金额。因此，在对18级4月消费者进行划分时，参考该模型，最终提取FM作为聚类指标。其中F为4月消费频次，M为4月消费金额。

在对M的指标进行选定时，考虑到18级学生中周末不消费的情况比较常见（共有430位学生在周末消费均值为0元，占总消费人数的13.42%），因此将对消费金额按工作日与非工作日进行划分，并进行统计分析辅助判断。最终获得 4月总消费次数、均笔消费金额、工作日均笔消费金额和非工作日消费金额的统计分析如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Count | Mean | Work\_Mean | Holiday\_Mean |
| count | 3203 | 3203 | 3203 | 3203 |
| mean | 45.76772 | 6.882285 | 6.839579 | 5.83207 |
| std | 21.85263 | 9.261938 | 9.346527 | 3.788335 |
| min | 1 | 0.2 | 0 | 0 |
| 25% | 30 | 5.12 | 5.02 | 4.035 |
| 50% | 46 | 6.11 | 6.02 | 5.9 |
| 75% | 61 | 7.34 | 7.27 | 7.66 |
| max | 128 | 300 | 300 | 66.87 |

表1 对四个变量的统计分析

从上表可见，工作日均笔消费金额和非工作日的均笔消费金额区别明显：**极差不同，且值的分布不同**，因此，在聚类分析中可以利用这两个指标进行聚类分析。

因此，在不同消费类型族群划分中，选择**月总消费频次、工作日均笔消费金额和非工作日均笔消费金额**进行划分。

3.2、利用聚类分析确定划分的定量标准：

在本次分析中，将利用KMeans算法进行聚类分析，初步确定分类标准。

在进行KMeans聚类分析前，需要对数据进行异常值和标准化处理。因此，需分别考察三个指标的数据分布情况，结果如下图显示：

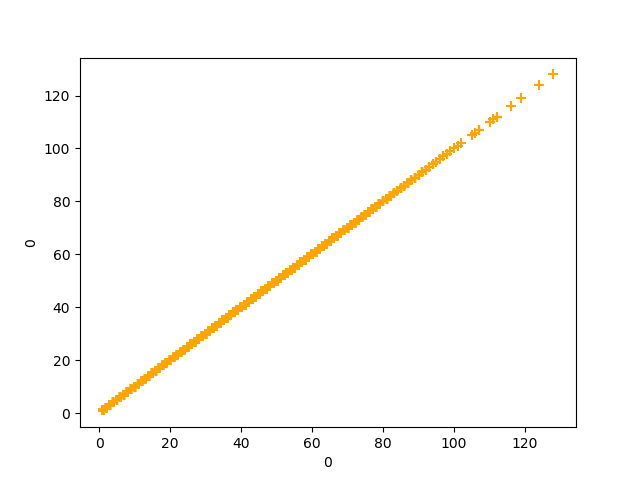


图3-1 月总消费频次数据分布

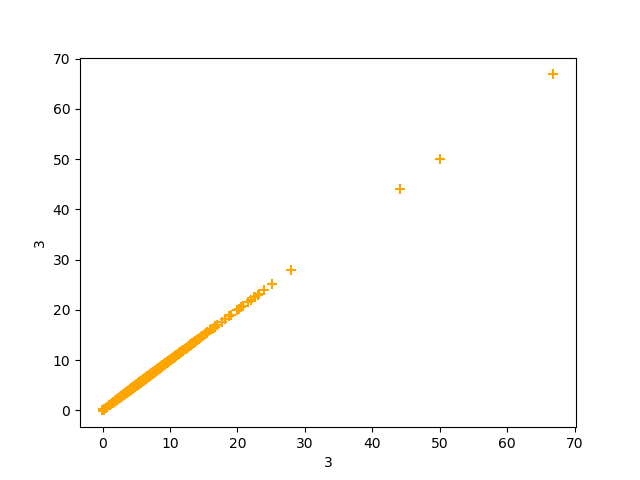
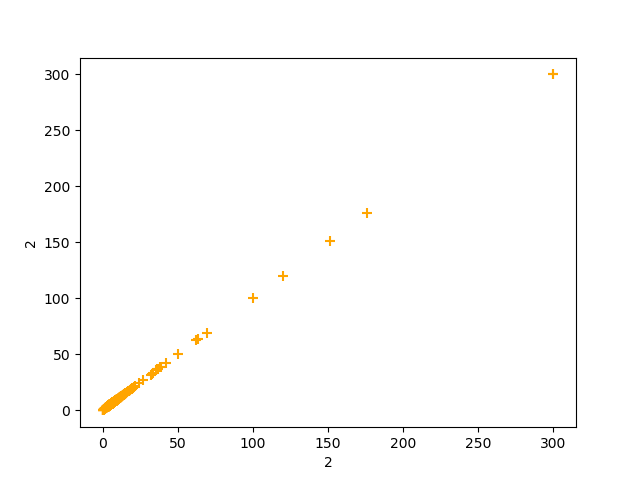


图3-2 工作日均笔消费金额数据分布 图3-3 非工作日均笔消费金额数据分布

由上图可见，仅工作日和非工作日均笔消费金额存在一些较大的值，且这些值都可以明显看出是远高于平均消费水平的。因此在进行聚类分析时，只对 **值小于22 的工作日均笔消费金额数据** 和 **值小于28 的非工作日均笔消费金额数据**进行聚类分析。

在经过异常值处理、离差标准化处理后，对数据集针对不同指标分别进行聚类分析，得到以下结果：

**月总消费频次：**当族数为2时，该指标的聚类评估效果最好，但是此时数据被分为[1, 45]和[46, 128]的两类，没有什么实际的效用。在考虑分成3类时，数据被分为：[1, 33], [34, 60]和[61, 128]，此时分类有意义。而在后续分析中，发现值在[54, 60]之间的消费者几乎在清明假期三天未在学校消费，考虑到清明节是一个饱受关注，举国活动的传统节日，因此，最后决定以 54次 做临界线 划分月总消费频次。

最终确定分组为[1, 53]和[54, 128]两组，**前者代表较少的消费频次，后者代表较多的消费频次**。

**工作日均笔消费金额：**当族数为2时，该指标的聚类评估效果最好，但是没有实际的效用。因此考虑分成3类。此时，数据被分为：[0, 6.13], [6.14, 9.9]和[9.96, 300]，结合前面对该指标的统计分析，确认该分类有意义。

最终确定分组为[0, 6.13], [6.14, 9.9]和[9.96, 300]三组，分别表示为**均笔消费低，均笔消费正常和均笔消费高**。

**非工作日均笔消费金额：**当族数为6时，该指标的聚类评估效果最好。分别为[0, 2.05], [2.1, 5.12], [5.13, 7.37],[7.38, 10.28],[10.32, 15.5]和[16, 66.87]。但是由于分类数量较多，不利于后续分析。因此，结合非工作日存在不消费的情况，考虑分成4类：**均笔消费极低或不消费**，**均笔消费低，均笔消费正常和均笔消费高**。

最终，数据被分为：[0, 2.05], [2.1, 5.63], [5.64, 10.28]和[10.32, 66.87]。

3.3、按不同消费特点进行族群划分

利用以上确定指标进行消费族群划分，描述其消费特点，并从中考虑给学校判定学生的经济情况提供参考意见。

# 四、分析结果

（一）、食堂数据分析及建议

### 1、早中晚各食堂就餐人次占比饼图结果分析：

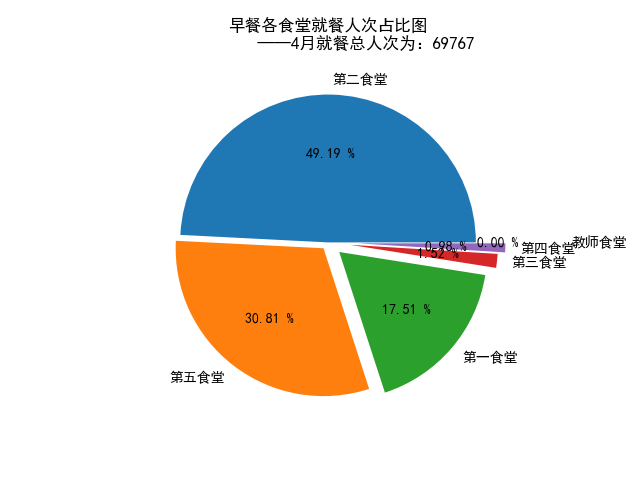


图4 早餐各食堂就餐人次占比图

从收集的数据中，总共有**69767条记录**为早餐段食堂消费记录。其中该时段各食堂就餐人次的排名为：

第二食堂 > 第五食堂 > 第一食堂 > 第三食堂 > 第四食堂 > 教师食堂

（34319） （21492） （12216） （1057） （682） （1）

其中，早餐时段几乎一半的人次都在第二食堂就餐，而第三、四和教师食堂就餐人次占比较小，特别是教室食堂4月份早餐段就餐人次只有一条记录。

可见，对于早餐时段的就餐选择而言，第二食堂是热门的就餐地点，而第三、四和教师食堂是冷门的就餐地点。

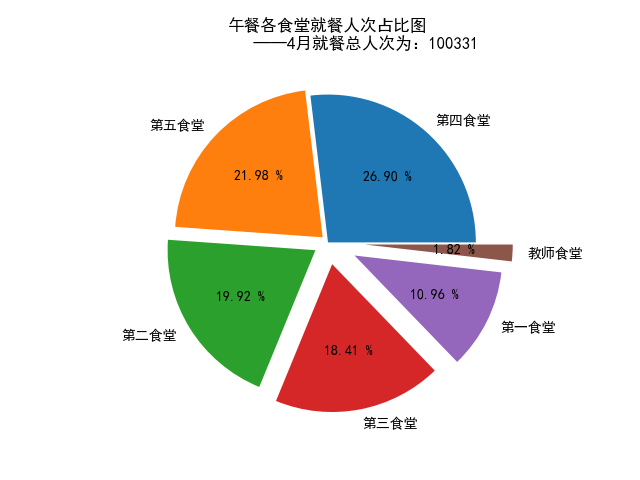


图5 午餐各食堂就餐人次占比图

从收集的数据中，总共有**100331条记录**为午餐段食堂消费记录。其中该时段各食堂就餐人次的排名为：

第四食堂 > 第五食堂 > 第二食堂 > 第三食堂 > 第一食堂 > 教师食堂

（26985） （22052） （19989） （18474） （11000） （1831）

从数据上看，午餐时段是校园食堂消费的高峰，4月就餐总人次是三个时间段里最高的。而此时各食堂就餐占比的排名发生了明显的变化，原来早餐时段就餐人次极少的第四食堂 成为中午最热门的食堂，而早餐最热门的就餐地—第二食堂 月总消费记录人次少了至少10000人次。但是，午餐时段没有出现一家食堂极其热门的情况，第四、五、二和三食堂起到较好的分流效果，占比差距不大。此外，教师食堂的就餐记录有很大的增加，且按日均下来，有63人/天 的消费人次，说明人们更愿意在中午来教师食堂就餐且来就餐人次不少。

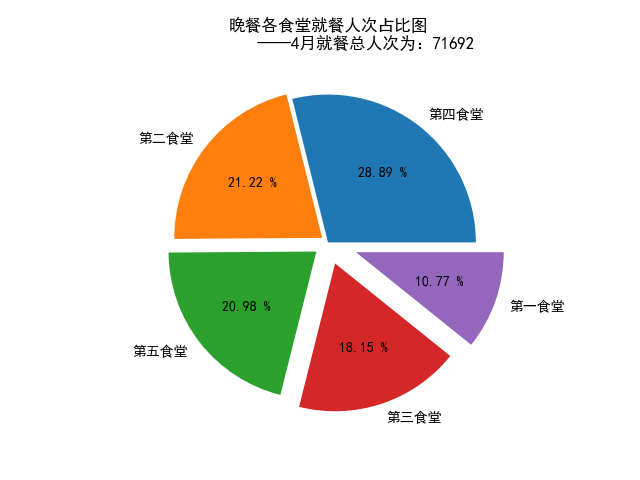


图6 晚餐各食堂就餐人次占比图

从收集的数据中，总共有**71692条记录**为晚餐段食堂消费记录。其中该时段各食堂就餐人次的排名为：

第四食堂 > 第二食堂 > 第五食堂 > 第三食堂 > 第一食堂

（20709） （15216） （15038） （13009） （7720）

从数据上看，晚餐时段就餐总人次是三餐中第二高的。在各食堂就餐人次占比上变化不大，仅是第二、五食堂交换了排名，但是教师食堂未出现在晚餐的记录中。此外，晚餐时段的各食堂就餐占更为均匀，其中第二、五和三食堂占比相差不大。

·由上可见，早中晚三个就餐时段的就餐人次排名有较为明显的变化。作为中午与晚餐的最热门就餐地点：第四食堂，在早餐就餐时段日均消费人次为24人/日，占比不到1%。而作为早餐最热门的就餐地点：第二食堂，在午晚餐就餐时段的就餐记录是逐步大幅度减少；而就教师食堂而言，仅有在午餐时段的就餐记录较多，在早餐时段4月仅有1条就餐记录，而晚餐时段更是没有就餐记录。

### 2、就占比饼图给学校食堂运营的建议：

从数据上看，学校食堂的消费记录不差，甚至午餐时段就餐记录能够达到3460条/天。

针对早餐时段，从饼图中可以看到在该时段第二食堂的就餐人次极大，说明此时该食堂的人流较密集，同时考虑到中午和晚上时段各食堂占比差距没有如此大，**建议学校考虑如何分流早餐时段第二食堂的人流到其他食堂**。其次，在所有的记录中，教师食堂的记录尤为特别，仅有午餐时段有相对最少的就餐人次，而早餐时段4月仅有一次记录，甚至晚餐时段没有记录。**建议学校考虑仅在午餐时段开放教师食堂，或针对这样的情况考虑更好的营销策略，为改善该食堂的运营情况**。

### 3、工作日与非工作日就餐曲线图：

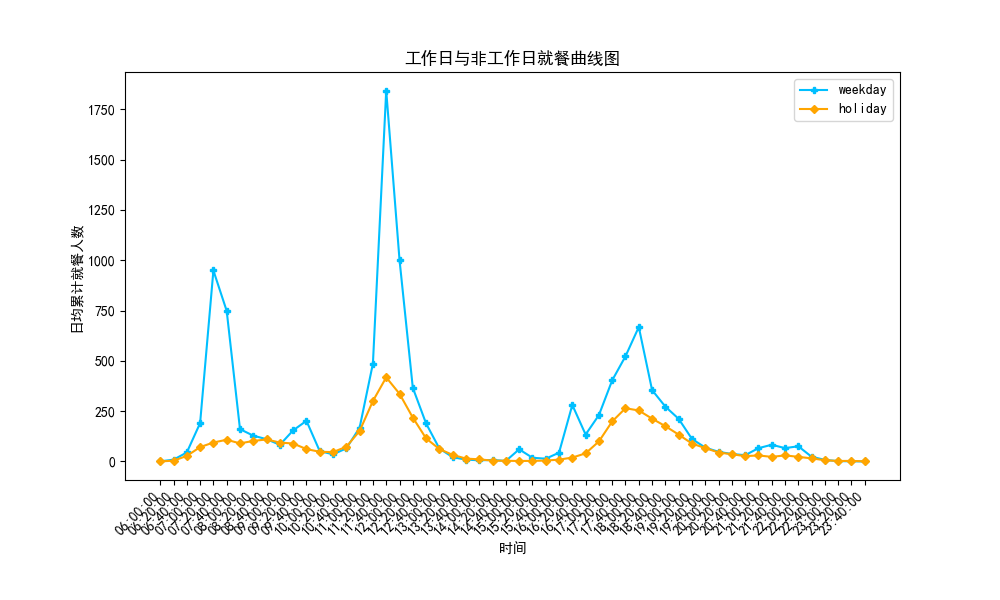


图7 工作日与非工作日就餐曲线图

由上图可见，在工作日一共出现了7个就餐峰段。分别为早餐时段的7：20，9：40；午餐时段的11：40；下午茶时段的15：00，16：20；晚餐时段的18：00；夜宵时段的21：20-22：00。

而非工作日一共出现3个就餐峰段。分别是早餐时段的7：40-8：40，午餐时段的11：40，和晚餐时段的17：40。

对**工作日的就餐峰段进行分析**：

1、早中晚餐时段，日均累计就餐人次突出。分别为早上的7：20，中午的11：40和晚上的18：00，都比较符和大学生的课程安排，都为课前，课后的时间段。

2、最明显的三个就餐峰段：早晚餐就餐峰段，以日均就餐人次超过1800最为突出。而另外两个峰段日均消费人次也超过500人。

3、对于9：40，15：00，16：20和21：20-22：00的就餐峰段，可以得出，在工作日除了日常三餐消费外，人们也会参与到其他消费中，说明人们在学校食堂消费的目的具有多样性。

对**非工作日的就餐峰段进行分析**：

1、仅有早中晚餐三个明显就餐峰段。其中，午餐就餐峰段仍旧和工作日就餐峰段相同，但早餐就餐峰段较工作日晚且延续时间长，而晚餐就餐峰段较工作日早。可见在非工作日，学生在学校消费的目的较为单一，且消费时间较工作日集中。

2、同样，其中以午餐时段的日均就餐人次最为突出，但少于500人次/日，甚至比工作日早餐与晚餐时段峰值更低。

3、在下午时段和晚上十点过后，日均消费人次接近为0.

### 4、以就餐时间曲线给学校食堂运营的建议：

1、从就餐时间曲线图可见，无论是工作日和非工作日，午餐时段的就餐人次是最多的，因此建议学校饭堂重视午餐时段的人手安排，以缓解较大人流造成的麻烦。

2、由于，非工作日仅有早午晚餐时段峰值明显，且在下午时段和晚上十点过后消费人次几乎为0，因此建议考虑减少非工作日除早午晚餐时段外的营业范围或甚至可以考虑不营业，以减少运营成本。

3、考虑到非工作日就餐总人次明显少于工作日就餐总人次，因此建议非工作日食材准备可以较工作日少。

（二）、18级学生消费数据分析及建议

### 1、 对整体消费行为进行分析

经统计，4月人均刷卡频次为：46次

人均消费为：285.18元

### 2、对不同专业不同性别族群进行分析

根据不同专业不同性别，计算人均消费频次及金额，如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 专业 | 人均消费频次（女） | 人均消费频次（男） | 人均消费金额（女） | 人均消费金额（男） |
| 18产品艺术 | 46 | 39 | 290.58 | 307.13 |
| 18会计 | 44 | 39 | 260.24 | 283.73 |
| 18动漫设计 | 50 | 42 | 319.93 | 276.7 |
| 18商务日语 | 46 | 45 | 303.44 | 320.78 |
| 18商务英语 | 44 | 70 | 271.02 | 399.34 |
| 18国贸实务 | 43 | 48 | 253.56 | 286.6 |
| 18国际商务 | 48 | 31 | 274.81 | 187.83 |
| 18国际金融 | 38 | 27 | 226.33 | 218.45 |
| 18宝玉石鉴定 | 42 | 30 | 267.44 | 250.62 |
| 18审计 | 49 | 48 | 250.63 | 280.99 |
| 18嵌入式技术 | 33 | 38 | 204.76 | 293.66 |
| 18工业工程 | 49 | 36 | 315.6 | 293.91 |
| 18工业机器人 | 50 | 55 | 291.36 | 343.62 |
| 18工业设计 | 56 | 41 | 324.37 | 260.05 |
| 18工商企管 | 55 | 37 | 276.31 | 210.57 |
| 18工程造价 | 53 | 36 | 304.25 | 251.55 |
| 18市场营销 | 43 | 32 | 240.94 | 228.51 |
| 18市政工程 | 50 | 37 | 295.6 | 276.45 |
| 18建筑工程 | 53 | 42 | 285.46 | 280.24 |
| 18建筑设计 | 58 | 52 | 298.09 | 317.51 |
| 18投资与理财 | 48 | 41 | 287.47 | 314.55 |
| 18旅游管理 | 45 | 46 | 274.09 | 367.65 |
| 18机械制造 | 58 | 52 | 330.21 | 325.41 |
| 18机械制造（学徒） | 61 | 33 | 447.46 | 201.45 |
| 18模具设计 | 62 | 53 | 401.89 | 340.91 |
| 18物流管理 | 58 | 52 | 301.68 | 298.22 |
| 18环境艺术 | 56 | 38 | 276.6 | 309.11 |
| 18电子商务 | 52 | 55 | 263.56 | 323.69 |
| 18电气自动化 | 41 | 59 | 302.95 | 361.74 |
| 18皮具艺术 | 43 | 51 | 250.58 | 334.49 |
| 18社会工作 | 41 | 44 | 253.48 | 307.85 |
| 18艺术设计 | 34 | 29 | 206.73 | 195.67 |
| 18视觉传播 | 48 | 55 | 324.39 | 392.64 |
| 18计算机应用 | 39 | 50 | 213.72 | 344.79 |
| 18计算机网络 | 38 | 43 | 211.39 | 326.92 |
| 18软件技术 | 34 | 37 | 238.04 | 268.82 |
| 18连锁经营 | 53 | 54 | 275.63 | 339.62 |
| 18酒店管理 | 36 | 47 | 229.17 | 405.08 |
| 18金融管理 | 42 | 30 | 252.78 | 230.76 |
| 18首饰设计 | 37 | 39 | 311.87 | 274.45 |

表2 不同专业不同性别消费特点

其中，根据月总消费频次指标，同一专业不同性别的极差出现在： 18机械制造（学徒）专业，对比柱状图如下图所示：

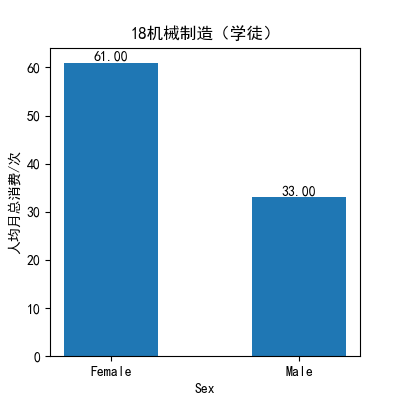
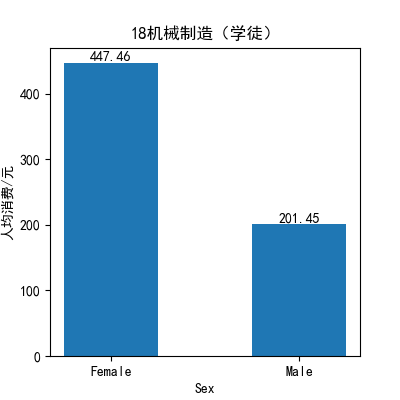
 

图8 18机械制造（学徒）男女性月人均总消费频次和金额图

其中女性的人均消费频次为61次，男性的人均消费频次为33次。说明，**该专业女性日常较依赖于学校部门进行日常消费，而男性则相反。**也因此，导致两者间人均月总消费金额差值是所有专业中最大的，为 447.46- 201.45 = 246.01 元。

此外，同一专业男性比女性消费金额的极差出现在：18酒店管理专业。对比柱状图如下所示：

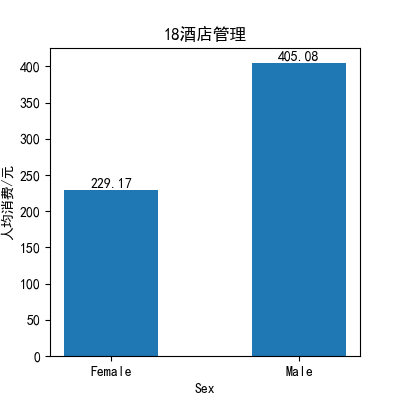
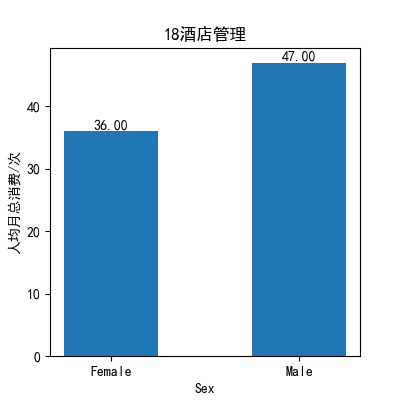
 

图9 18酒店管理男女人均消费金额和月总人均消费频次图

其中，不同性别的月总人均消费金额差值的为 405.08 - 229.17 = 175.91元。

但是从月总人均消费频次的差值来看相差不大，可见月总消费金额极差的出现主要原因为 人均消费金额/月总消费频次 两者差距较大。经计算，男性为 8.62元/次，为所有专业中最高的；而女性为6.27元/次，略高于均值。说明，**该专业男同学较女同学在校园消费单笔消费金额更多高，该专业男同学乐意在校园超额消费**（对比单笔消费金额均值6.1）。

### 3、对不同消费类型族群进行分析

根据三个聚类指标，最终分类结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 值范围 | 个数 | 标记 |
| 较少消费频次 | 1-53 | 1998 | Label\_0 =0 |
| 较多消费频次 | 54-128 | 1205 | Label\_0 =2 |

表3 月总消费频次聚类结果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 值范围 | 个数 | 标记 |
| 较低水平消费金额 | 0-6.13 | 1691 | Label\_2 =0 |
| 正常水平消费金额 | 6.14-9.9 | 1320 | Label\_2 =1 |
| 较高水平消费金额 | 9.96-300 | 192 | Label\_2 =2 |

表4 工作日均笔消费金额聚类结果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 值范围 | 个数 | 标记 |
| 极低消费或不消费 | 0-2.05 | 481 | Label\_3 =0 |
| 较低水平消费金额 | 2.1-5.63 | 977 | Label\_3 =1 |
| 正常水平消费金额 | 5.64-10.28 | 1515 | Label\_3 =2 |
| 较高水平消费金额 | 10.32-68.87 | 690 | Label\_3 =3 |

表5 非工作日均笔消费金额聚类结果

将所有数据按以上2\*3\*4=24类标签进行分组。最终得到23组不同消费类型的族群。对每个族群的个数及特征描述如下表所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 标签 | 数量 | 频次 | W\_M | H\_M | 特点 |
| 000 | 202 | 22.50 | 4.36 | 0.21 | 消费频次极低，工作日均笔消费金额低，非工作日几乎不在校园消费。 |
| 001 | 355 | 37.35 | 4.77 | 4.26 | 消费频次低，工作日和非工作日均笔消费金额低 |
| 002 | 329 | 37.40 | 5.17 | 7.23 | 消费频次低，工作日均笔消费金额低，但非工作日均笔消费金额高于正常水平 |
| 003 | 40 | 29.48 | 5.26 | 14.02 | 消费频次极低，工作日均笔消费金额在正常水平，但非工作日几乎不消费 |
| 010 | 199 | 22.24 | 7.61 | 0.10 | 消费频次极低，工作日均笔消费金额在正常水平，但非工作日几乎不消费 |
| 011 | 161 | 35.27 | 7.24 | 4.34 | 消费频次低，工作日均笔消费在正常水平，但非工作日均笔消费金额较低 |
| 012 | 440 | 36.29 | 7.46 | 7.68 | 消费频次低，但工作日与非工作日均笔消费金额在正常水平 |
| 013 | 102 | 32.02 | 7.71 | 13.49 | 消费频次低，工作日均笔消费金额正常，非工作日均笔消费金额高于正常水平 |
| 020 | 54 | 13.44 | 36.73 | 0.03 | 消费频次极低，工作日均笔消费金额超常，非工作日几乎不消费 |
| 021 | 21 | 29.52 | 16.43 | 4.39 | 消费频次低，工作日均笔消费金额超常，但是非工作日均笔消费金额低 |
| 022 | 57 | 28.51 | 14.10 | 7.94 | 消费频次低，非工作日均笔消费金额在正常水平，工作日均笔消费金额高 |
| 023 | 38 | 28.39 | 12.76 | 15.06 | 消费频次低，但工作日与非工作日均笔消费金额都高于正常水平 |
| 100 | 14 | 58.50 | 4.89 | 0.91 | 消费频次稍高，但工作日均笔消费金额低，非工作日几乎不消费 |
| 101 | 375 | 69.63 | 4.56 | 4.44 | 消费频次高，但工作日与非工作日均笔消费金额低 |
| 102 | 363 | 68.31 | 5.28 | 6.86 | 消费频次高，工作日均笔消费金额低，但非工作日均笔消费金额于正常水平 |
| 103 | 13 | 65.85 | 5.54 | 11.81 | 消费频次稍高，工作日均笔消费金额低，但非工作日均笔消费金额高 |
| 110 | 12 | 58.50 | 6.97 | 0.13 | 消费频次稍高，工作日均笔消费金额于正常水平，但周末几乎不消费 |
| 111 | 64 | 68.66 | 7.11 | 4.83 | 消费频次高，工作日均笔消费金额正常，但周末均笔消费金额低 |
| 112 | 312 | 66.80 | 7.15 | 7.58 | 消费频次高，工作日与非工作日均笔消费金额正常 |
| 113 | 30 | 65.33 | 7.74 | 11.72 | 消费频次高，工作日均笔消费金额正常，非工作日均笔消费金额较高 |
| 120 | / | / | / | / | / |
| 121 | 1 | 55.00 | 14.12 | 4.70 | 消费频次刚过水平线，工作日均笔消费金额高，非工作日均笔消费金额低 |
| 122 | 14 | 67.00 | 11.38 | 7.24 | 消费频次高，工作日均笔消费金额稍高，非工作日均笔消费金额在正常水平 |
| 123 | 7 | 63.86 | 15.77 | 19.38 | 消费频次高，同时工作日与非工作日均笔消费金额都高 |

表6 聚类结果展示

### 4、给学校判定学生的经济状况的参考意见：

从以上的不同消费特征族群分类中，**类201**值得留意。它消费频次高，说明该类消费者日常生活中依赖于学校部门的服务。工作日与非工作日均笔消费金额都较低且两者差值不大，说明假期不影响其在学校进行消费，且每次的消费都比较克制，仅仅为了刚好满足日常需求。而这样的消费族群，与需要学校经济补助的族群具有相似的消费特征。**因此对于 月总消费频次大于等于54次，工作日均笔消费不高于6.13元，且非工作日均笔消费不高于5.63元的学生，学校可以将其作为需要经济补助的参考对象。**