# 学生校园校园行为分析报告

惠州学院 16应用统计学1班 叶宇轩

任务1

数据发下来后，可以得知data1.csv为学生ID表、data2 cs为学生的消费记录表。所以我使用JetBrains PyCharm导入data1.csv、 data2.csv并完成以下任务：

任务1.1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。

任务1.2对学生个人信息表和消费记录表进行关联。

我在这里数据预处理的方法有：在删除无用列，删除对研究本问题无关联的列，让数据更具清晰性；检测和处理缺失值,删除或者替换这些缺失值；去除重复值，重复值会影响数据得分方差和数据的准确性；检测然后去除异常值，但要注意数据的分布是否适合检测的方法。表关联的依据或者说主键使用CardNo，因为这代表校园卡号，而每位学生的校园卡号都唯一。如果由于工作人员的疏忽，忘记而缺失或者由于数据采集器等故障等原因造成缺失的缺失值，我可以用这个方法data.isnull().sum()把缺失值检测出来并直接删除。异常值处理，对异常的数据进行更改或填补，我选用的方法是直接删除，简单快捷，因为样本容量的关系也不太影响数据输出的准确性。

异常日期，也就是节假日的处理可以用loc方法将数据删除或者如果需要的话就替换。

直接两张表的合并的话用merge方法就可以了。最后把表导出用data\_mer.to\_csv(' data\_mer.csv',encoding=’gbk’)，不然会出现乱码。

任务2

对食堂就餐行为分析，井完成下列操作。

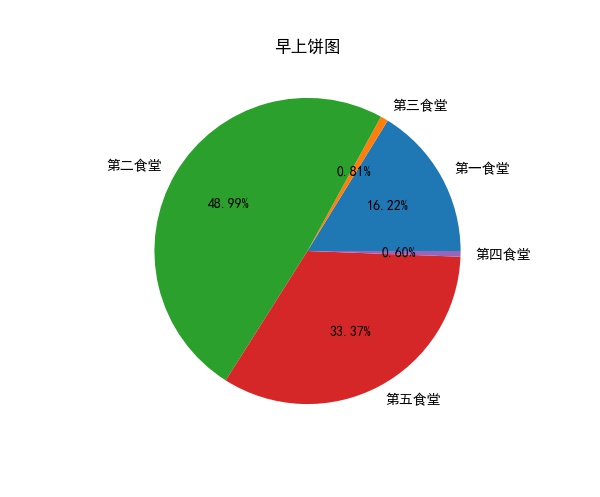
任务2.1绘制各食堂就餐人次的占比饼图，分析学生早中晚餐的就餐地点是否有显著差别。(提示: 时间间隔非常接近的多次刷卡记录可能为一次就餐行为)

既然是求早中晚餐的就餐地方，那么就要将时间数据分成三个时间段了。同时就餐人次在同一地点、同一时间范围内的消费记录，都将看作一次就餐行为。那么这就需要python的代码分步骤去完成了。

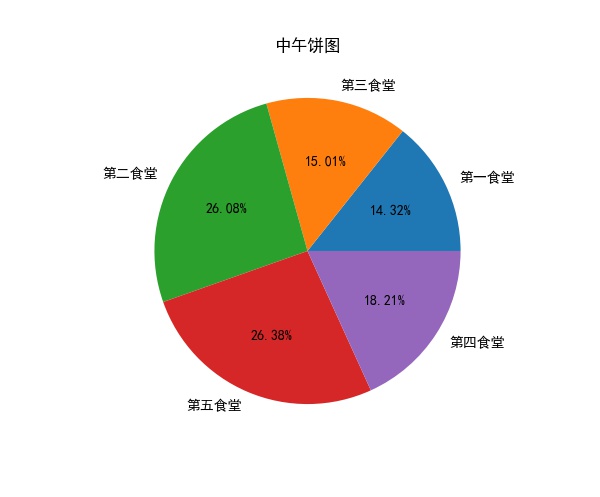
首先要对任务1中清洗完数据且合并后的表中的Date进行操作，用pd.to\_datetime方法进行转换，然后用循环分别进行年月日、星期甚至是时的提取，提取出来的列可以到已经在原本的表格右边中生成了，这时候操作就很方面了，不用对原本繁杂的日期数据进行操作。右边新增的列中，从1到7的weekday星期和和1到24的hour小时是对于解决任务至关重要的。然后就是划分早中午了，譬如，以中午为例（其实我的划分也包含的下午），data\_afternoon = data3.loc[(data3['hour'].apply(lambda x: x in [11, 12, 13, 14,15])), :]这行代码直接选出从11点到15点的时间的行。然后进行分组统计，分组计算就很方便了。一下是结果：

早上：

​早餐地点人数排名：第二食堂>第五食堂>第一食堂>第三食堂>第四食堂

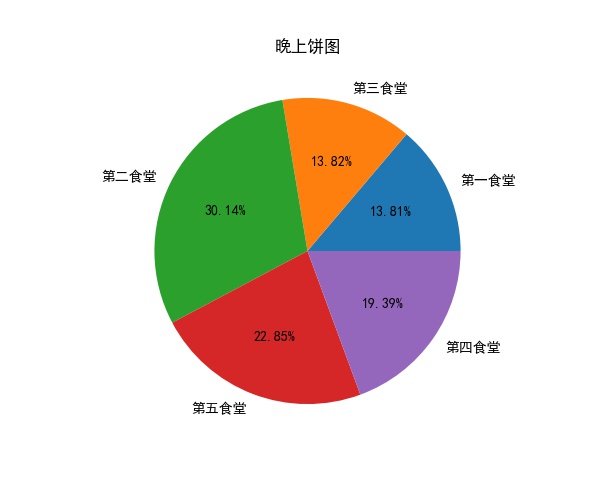


中午：



中午地点人数排名：第五食堂>第二食堂>第四食堂>第三食堂>第一食堂

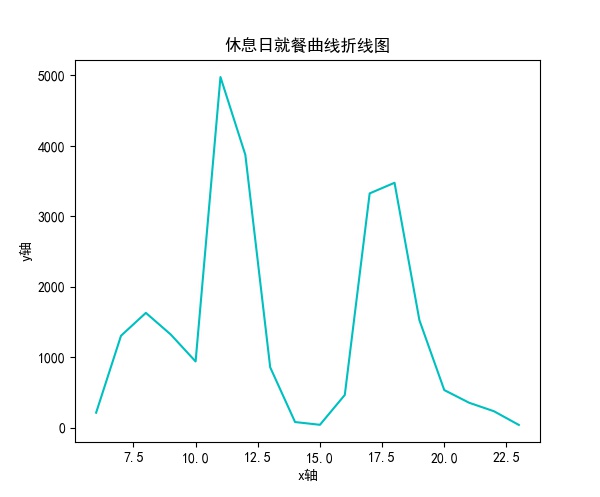
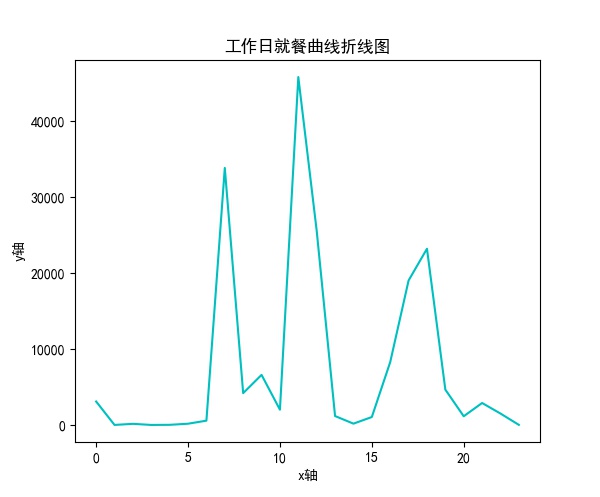
晚上



晚上地点人数排名：第二食堂>第五食堂>第四食堂>第三食堂>第一食堂

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

根据题目的意思，要将数据分成工作日和非工作日两个时间段，方法和任务2.1大致相同，将weekday提取出来，然后用plt画图并导出。



任务3

任务3:对学生消费行为分析，并完成下列操作。

任务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，展示某一专业不同性别学生群体人均消费对比柱形图

计算整体和不同专业间的，用到分组求和和分组平均值。在这里我选用下面两条语音进行计算：

分组求和data.groupby(data4['Money']).sum()

分组求平均值data\_mean= data4['CardNo'].groupby(data4['Money']).agg({sum,np.mean})

结果在任务3作业中已上传。

任务3.2 展示某一专业不同性别学生群体人均收入

这里选取18国际金融专业，通过分组求和用isin筛选的方法进行分组求和，得出结果：

任务3.3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，并确定聚类特征。

在这里，我选用的聚类模型是RFM模型，即：

(1) R (Rencency) :最近一次消费

(2) F(Frequency) :消费频率。

(3) M (Monetary) :消费金额。

因为样本容量大，这里我选择以点位基准的KMeans法，分成三簇，同时kmeans\_model.labels\_和kmeans\_model.cluster\_centers这两个方法进行聚类模型的查看，并用plt.scatter可视化来检测拟合的优度。