项 目 分 析 报 告

百货商场用户画像描绘与价值分析

甲方：百货商场

项目参与人姓名：温 宇

项目提交时间：2020.2.29

**目录**

[1 项目介绍 3](#_Toc33912486)

[1.1 项目背景 3](#_Toc33912487)

[1.2 项目数据及目标介绍 3](#_Toc33912488)

[2 数据准备 4](#_Toc33912489)

[3 数据处理 5](#_Toc33912490)

[3.1 重复值处理 7](#_Toc33912491)

[3.2 缺失值处理 8](#_Toc33912492)

[3.3 增设变量 9](#_Toc33912493)

[3.4 筛选符合逻辑的数据 10](#_Toc33912494)

[3.5 会员处理 11](#_Toc33912495)

[4 统计与可视化分析 11](#_Toc33912496)

[4.1 统计分析 11](#_Toc33912497)

[4.2 可视化分析 13](#_Toc33912498)

[4.3 总结 21](#_Toc33912499)

[5 用户画像 22](#_Toc33912500)

[5.1 构建会员基本特征标签 22](#_Toc33912501)

[5.2 构建会员业务特征标签 23](#_Toc33912502)

[5.3 构建会员偏好特征标签 24](#_Toc33912503)

[5.4 建立用户画像 24](#_Toc33912504)

[6 用户精细化分和K-means聚类 25](#_Toc33912505)

[6.1 基于LFM模型对用户进行化分 25](#_Toc33912506)

[6.2 基于K-means算法对用户进行聚类 26](#_Toc33912507)

[6.3 小结 27](#_Toc33912508)

# 项目介绍

**1.1 项目背景**

美国Target超市通过女生购买大量维生素补给和专用乳液等典型孕妇商品预测怀孕，并把母婴系列产品优惠券和信息混合在其他产品信息中发送给顾客，使得销量增加。除此之外，商场还通过打折促销、捆绑销售、会员积分等手段吸引顾客，从而增加利润。

本项目来自国内某百货商场，正面临老会员流失，新会员获取困难等经营危机，在巨大的竞争压力下，该商场经常使用打折促销来吸引消费者，但是没有指导性的打折促销活动不仅达不到商场预期的效果，反而增加了成本。

商场希望有针对性地实施营销策略来加强与会员地良好关系，比如，推出面向会员地营销活动，以此来维系会员的忠诚度，但是可能有人认为，维系老会员的成本太高。但事实上，发展新会员的资金投入远比维系现有会员成本要高。另一方面，会员价值不但体现在持续不断地为商场带来稳定的收入，同时也为商场制定营销策略提供数据支持。因此，完善会员用户画像，加强对现有会员的精细化管理，定期向其推送产品和服务，与会员建立稳定的关系，是该商场的得以更好发展的有效途径。

**1.2 项目数据及目标介绍**

我们通过甲方（百货商场）提供的两份数据——会员信息表和销售流水表。

第一份是商场会员信息表，包含4个字段，其数据字典如下表所示：

**表1 商场会员信息表数据字典**

|  |  |  |
| --- | --- | --- |
| **变量名** | **变量解释** | **变量类型** |
| kh | 会员卡号 | Numeric |
| csrq | 会员出生日期 | Datetime |
| xb | 会员性别 | 0 = 女, 1 = 男 |
| djsj | 会员入会时间 | Datetime |

注：数据于cumcm2018c1.xlsx中

第二份是商场的销售流水表，时间从2015年1月1日到2018年1月3日，共包含12个字段，其数据字典如下表所示：

**表2 商场销售流水表数据字典**

|  |  |  |
| --- | --- | --- |
| **变量名** | **变量解释** | **变量类型** |
| kh | 会员卡号 | Numeric |
| dtime | 订单产生时间 | Datetime |
| spbm | 商品编码 | String |
| sl | 销售数量 | Numeric |
| sj | 商品售价 | Numeric |
| je | 消费金额 | Numeric |
| spmc | 商品名称 | String |
| jf | 消费得到积分 | Numeric |
| syjh | 收银机号 | Numeric |
| djh | 单据号 | Numeric |
| gzbm | 柜组编码 | Numeric |
| gzmc | 柜组名称 | String |

注：数据于cumcm2018c2.csv中

本项目中甲方提供了两个明确目标：

**目标1：针对会员消费情况描绘用户画像**

用户画像是用户信息标签化，所谓标签，就是浓缩精炼的带有特殊含义的一系列词语，用于描述真实的用户自身带有的属性特征。如果用户信息足够丰富，可以可视化呈现用户画像结果，本项目，甲方希望我们通过数据分析技术详细分析会员的基本特征和消费情况，构建用户标签库，进而描绘用户画像。

**目标2：对会员进行精细化分，并针对不同群体制定对应的营销策略，从而提升销售利润**

这个目标与背景介绍中甲方面临的问题有着直接相关的关系，现有的促销活动缺乏指导性，不仅达不到预期的效果，还会增加成本，因此甲方希望对会员用户进行分类，并针对不同的用户群体指定对应的音效策略，从而提升销售利润。

# 数据准备

本次百货商场的数据有Xlsx和CSV两种格式，分别存放在cumcm2018c1.xlsx和cumcm2018c2.csv中

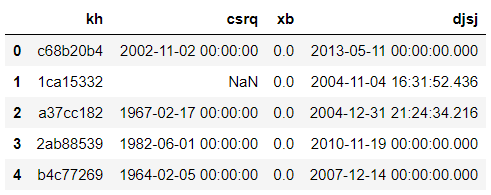
首先，分别导入（读取）文件名为cumcm2018c1.xlsx和cumcm2018c2.csv的数据。

|  |
| --- |
| import pandas as pd  c1=pd.read\_excel('C:/Users/50502/Desktop/Python项目/02-百货商场data/data/cumcm2018c1.xlsx')  c2=pd.read\_csv('C:/Users/50502/Desktop/Python项目/02-百货商场data/data/cumcm2018c2.csv') |

对表1查看导入的数据是否有误；

|  |
| --- |
| c1.head() # 查看前5行数据 |

运行可得如下结果：



对表2查看导入的数据是否有误；

|  |
| --- |
| c2.head() # 查看前5行数据 |

运行可得如下结果：



# 数据处理

数据清洗就是处理缺失数据以及清除毫无意义的信息。通过查看数据的各个属性来发现数据中存在的缺失值、异常值等，由于本项目数据集的缺失值不能简单进行填补，于是在合并数据表的过程中采用merge函数的默认属性，使用shape函数查看数据形状，使用info函数查看数据的非空值个数和数据类型，使用describe函数查看数据的描述统计信息，使用isnull函数和sum函数结合查看每列数据的空值。

|  |
| --- |
| # 合并数据集  c = pd.merge(c1,c2)  c.head()  # 查看数据的形状  print("数据的形状：", data.shape)  # 查看每列数据的非空值的个数和数据类型  print("每列数据的非空值的个数和数据类型：\n", data.info())  # 查看每列数据的描述统计信息  print("数据的描述统计信息：\n", data.describe())  # 查看每列数据的空值个数  print("每列数据的空值个数:\n", data.isnull().sum()) |

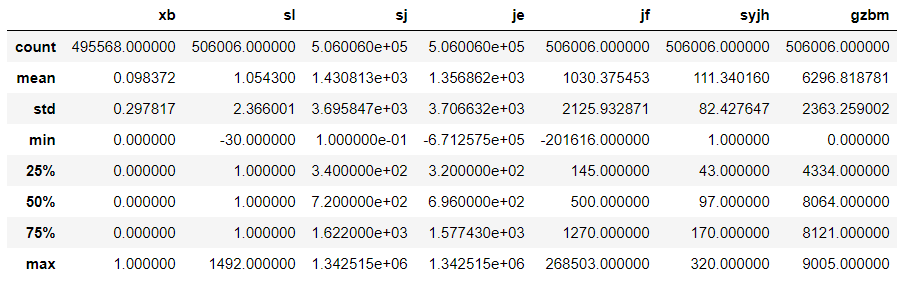
运行可得如下结果：

数据的形状：(506006, 15)

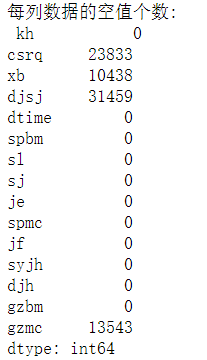
每列数据的非空值的个数和数据类型：



数据的描述统计信息：



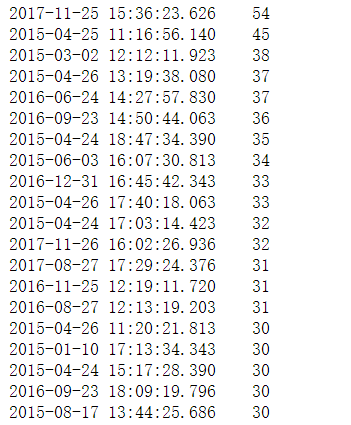
每列数据的空值个数:



使用value\_counts()函数可以查看数据中每个值的个数，以“销售日期”列为例：

|  |
| --- |
| # 查看销售日期列数据中每个值的个数  c.dtime.value\_counts() |

运行代码可得如下结果：



通过有观察以上属性值，发现同一时刻有许多相同的订单产生，推测有可能是由于重复值而产生，为此，接下来进行重复值处理，并通过3。

## 重复值处理

drop\_duplicates()：把数据结构中行相同的数据去除（保留其中一行）

|  |
| --- |
| # 数据去重  new\_c = c.drop\_duplicates()  print(new\_c.shape) |

运行代码可得到：

|  |
| --- |
| **(506006, 15)** |

由结果可看到，数据形状对比之前没有变化，说明尽管有多同一时刻产生多笔订单，但是并没有重复的情况。

## 缺失值处理

对于缺失值的处理方式，有数据补齐、删除对应行、不处理等方法。

dropna()：去除数据结构中值为空的数据行。

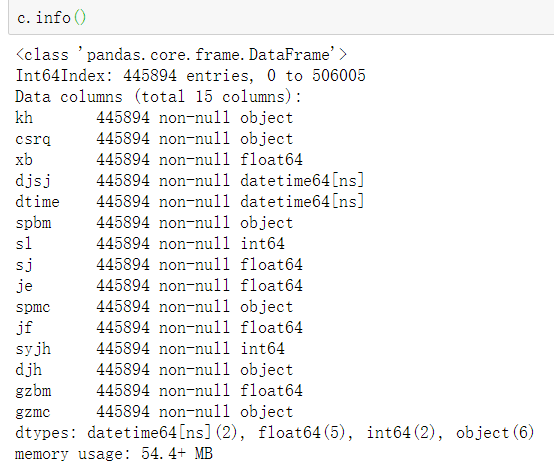
|  |
| --- |
| # 去除数据中的空值  c = new\_c.dropna()  c.shape |

运行代码可得到：

|  |
| --- |
| **(445894, 15)** |

可以看到数据形状由(506006, 15)变更为(445894, 15)，说明使用dropna()已经将数据中值为空的数据行去除了。

再次使用info函数查看数据非空值的数量，得到的运行结果如下图所示：



## 增设变量

为了更好绘制出会员的用户画像，通过已有的变量，处理成新的变量，便于体现用户画像，增设以下几个变量：年龄、年龄阶层、季节、会员入会时长，如下：

|  |
| --- |
| #定义年龄0-44为青年，45-59为中年人，60以上为老年人  #在c表中csrq  import datetime as dt  #提取年数  c['csrq']=pd.to\_datetime(c['csrq'],errors='coerce')  c['year'] = c['csrq'].dt.year  #年龄计算  now\_year = dt.datetime.now().year  c['age'] = now\_year - c['year']  #年龄阶层  bins = [0,44,59,120]  labels = ['青年 ','中年人', '老年人']  c['年龄阶层'] = pd.cut(c['age'], bins, labels=labels)  #季节  c['month']=c['dtime'].dt.month  bins = [0,2.9,5,8,11,12.1]  labels = ['冬季 ','春季', '夏季', '秋季','冬季']  #labels = ['Winter ','Spring', 'Summer', 'Autumn','Winter']  c['季节'] = pd.cut(c['month'], bins, labels=labels)  #入会时长数据定义、处理与描述  c['djsj']=pd.to\_datetime(c['djsj'],errors='coerce')  c['入会时长']=dt.datetime.now().year-c['djsj'].dt.year |

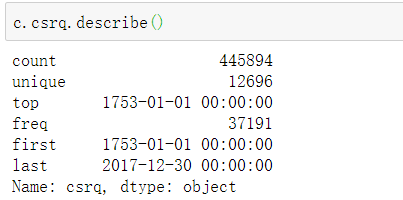
运行后进行head()查看前五行数据，得到结果如下：



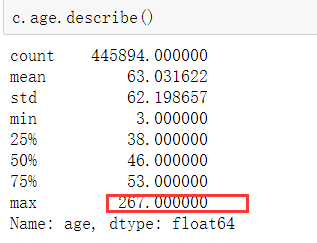
## 筛选符合逻辑的数据

通过describe函数观察会员的出生日期

运行结果如下：



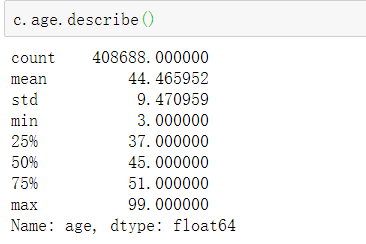
发现出现了1753这样不合理的数字，我进一步查看age变量，运行结果如下：



发现存在有267岁这样不合理的年龄，为了使数据更加合理，我设120岁为最大上限，代码如下：

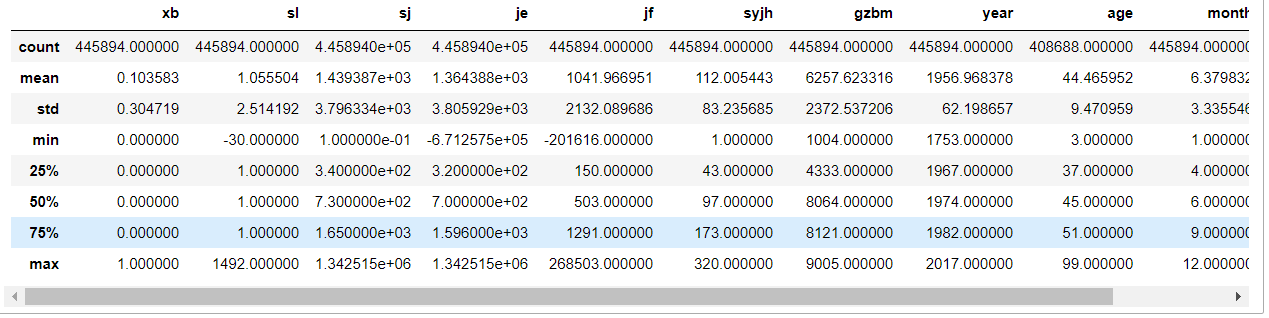


得到的结果运行如下：



根据数据的最大值为99，属于正常范围。

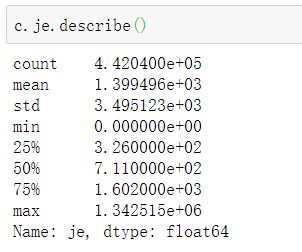
通过对整体数据用describe()函数进行观测，如下图所示：



发现消费金额存在不合理的值（金额为负数）为此，通过



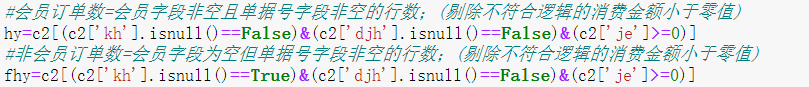
运行后得到的结果如下图所示：



至此完成基于逻辑正确的数据处理。

## 会员处理

根据甲方提供的数据，我们发现表1中的都是会员，而表二中存在非会员的订单，我们定位非会员为没有卡号的人，因为非会员只有订单没有卡号，为此，分别将会员和非会员存储于表格hy和fhy中，过程如下：



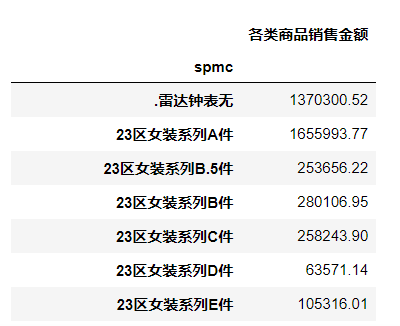
# 统计与可视化分析

## 统计分析

分组分析是根据分组字段将分析对象划分成不同的部分，以对比分析各组之间差异性的一种分析方法。这里我统计数据中商品的销售金额：

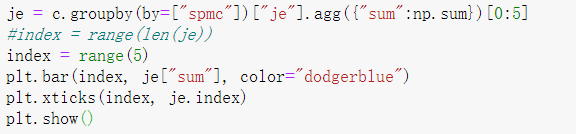
|  |
| --- |
| import numpy as np  c.groupby(by=["spmc"])["je"].agg({"各类商品销售金额":np.sum}) |

运行代码可得到：

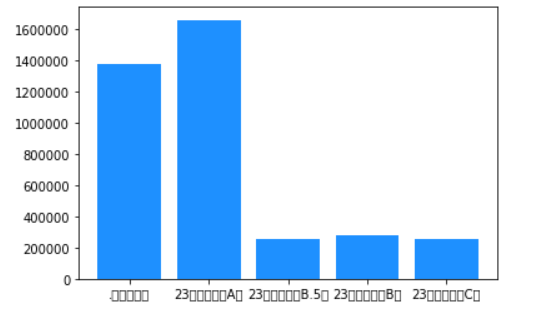




如图可以发现，有5936种商品，然而因为有许多不能明确种类的商品，故而这里没办法对所有的商品种类进行大类分类，其前五位用柱状图表示如下：



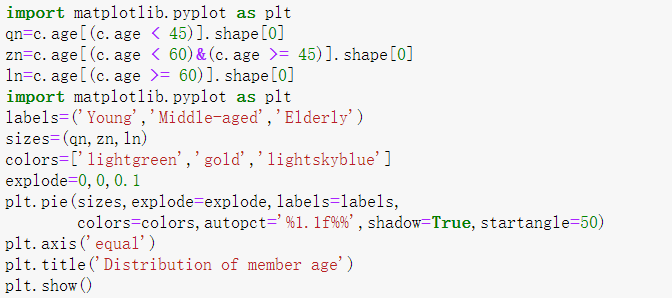
运行结果如下：



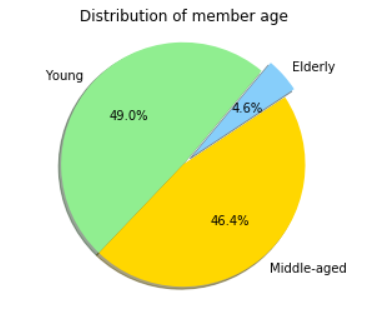
## 可视化分析

**4.2.1 会员年龄阶层占比**

为了直观展示会员的年龄阶层分布，通过matplotlib库调用pie函数用饼图的方式对年龄分布进行直观展示，过程和代码如下图所示：



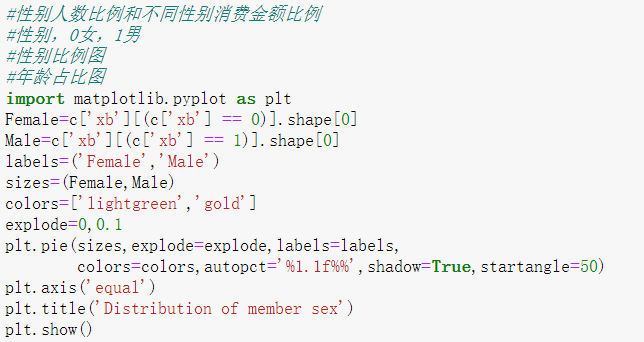
运行得到的结果如下图所示：



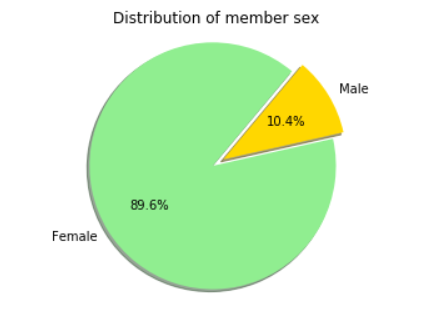
通过观察可以发现，该商场的会员大多数是青年人和中年人，占到95.4%，为此，商场在制定策略时，可以有倾向地把主要客户定位在青年人和中年人。

**4.2.2 会员性别占比**

众所周知，商场客户中不同性别地行为模式有着很大地区别，为了给商场进行定位和商品种类的占比，对会员的年龄占比进行分析，过程如下：



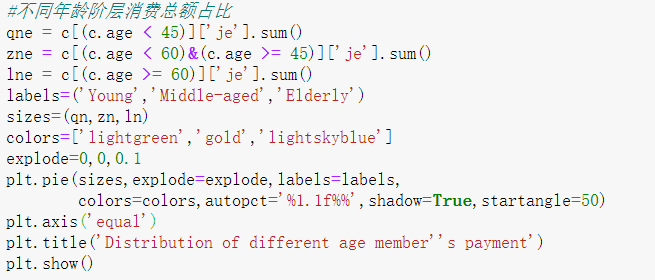
运行结果如下图所示：



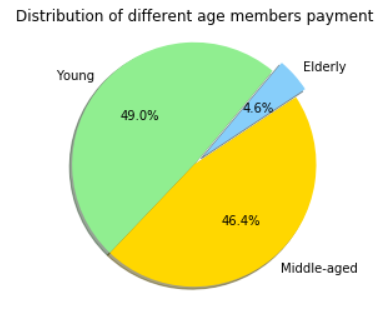
如图可以发现，该商场的89.6%主要顾客是女性顾客，因而在考虑商场商品的种类时，应当尽可能多地出售女性商品从而获得更高的利润。

**4.2.3 年龄阶层消费总额占比**

尽管有了会员年龄阶层占比和性别占比，但是，不同年龄阶层和性别尽管数量不大相同，然而金额也不一定会因此会有巨大差别，因为男性和中年人相对会有更多地金钱，为了探究这里面地关系，我绘制了不同年龄阶层会员地消费总额占比图，进一步观察，过程如下：



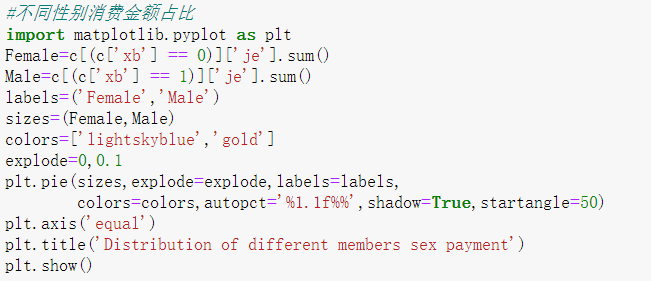
运行结果如下：



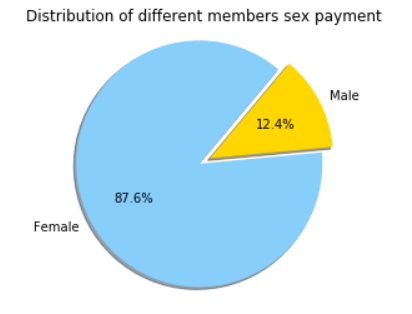
通过上图印证了之前的想法，尽管青年人和中年人会员占比相近，但是中年人的消费金额占比比青年人多，占比有53.2%。

**4.2.4 不同性别消费金额总数占比**

同样，因为不同性别的消费总额可能也会因此有差异，为此，我们绘制不同性别下的会员消费占比的图，过程如下：



运行结果如下：

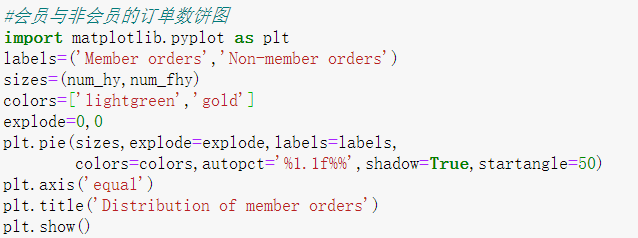


如图所示，相比不同性别会员数量的占比，消费金额占比女性稍微下降，原来为89.6%，如今为87.6%，可见，男性的消费金额就个体而言比女性要高，但总体达不到女性水平。

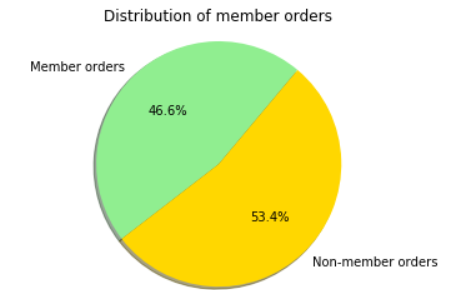
**4.2.5 会员与非会员的订单数量比**

为了分析来该商场购物的是会员多还是非会员多，他们的消费占比如何，考虑是否增加会员的数量，为此，首先了解会员于非会员的订单数量占比，过程如下：





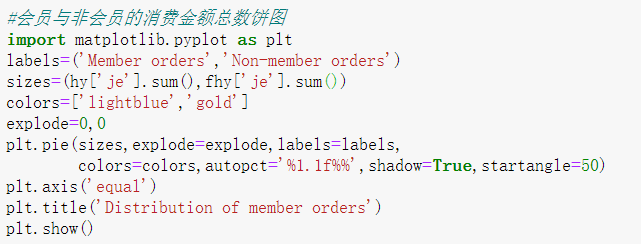
运行得到的结果如下图所示：



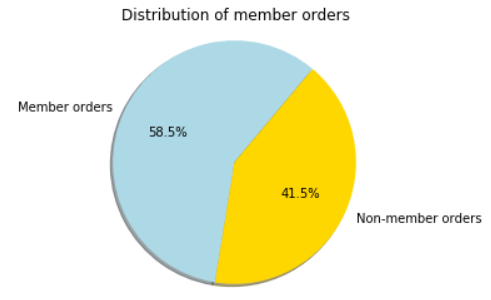
可以发现，就产生的订单数量来说，大多是由非会员产生而非会员产生。

**4.2.6 会员与非会员的消费金额占比**

我们有了会员与非会员在商场的订单产生数量占比，但是为了了解他们订单所产生的消费金额，我们进一步分析他们的消费金额占比，过程如下：



运行结果如下：



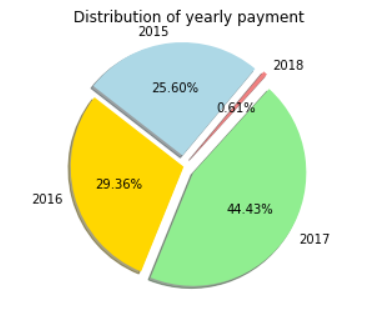
可以发现，消费金额来说，会员比非会员消费金额的总数要大，尽管非会员产生的订单消费记录比较多，因此，就商场的长远规划来说，应当尽可能发展多的会员。

**4.2.7 年度消费金额占比**

为了感受商场近些年的运营状况，用每年消费的总金额进行衡量，绘制年度消费的占比图，过程如下：



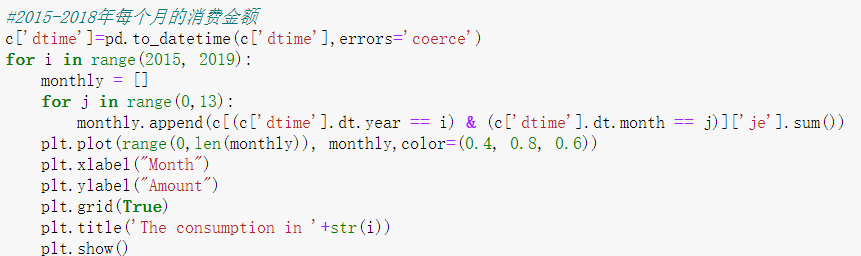
运行结果如下：



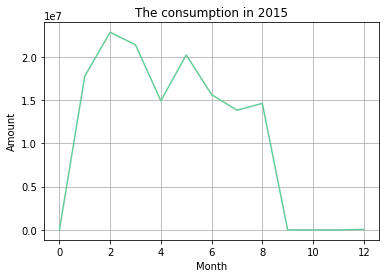
观察上图可以发现，除了2018年因为只有1个月的数据所以数值较小，该商场的会员每年的消费金额都在上升，可见发展会员是有助于商场长期的发展的。

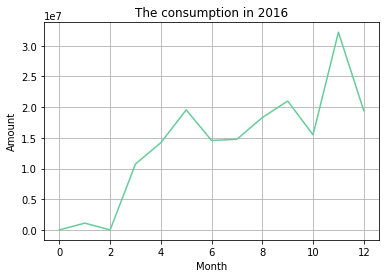
**4.2.8 2015-2018年每个月的消费金额**

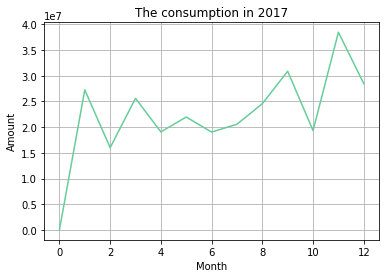
为了了解2015-2018年这些年各月的消费金额趋势，使用了折线图的表达方式，可以显而易见的表达出淡旺季，过程如下：

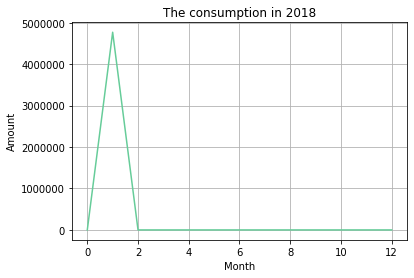


运行结果如下图所示：







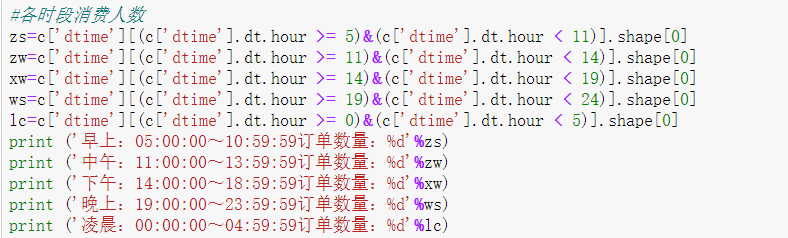


通过观察上图可以发现，除了2015年的数据较为反常，每年的2月份几乎都是每年中的历史最低点，因为正处过年时节，此时1月份会相对较高，因为大家需要购置年货，作为商家，可以尽量避免在2月份引进大量的商品，在1月或年末11-12月引进商品，避免导致滞销和过期。

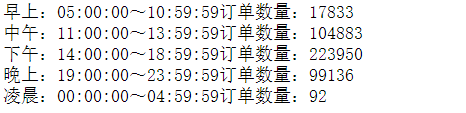
**4.2.9 一天中各时段的消费产生订单数量**

因为对于商场而言，人工通常是按时间段和小时计算的，不同时间段的工作人员数量的设置不同影响着商场的运营成本，为此，特别统计一天中的多个时段，研究不同时段的消费金额，过程如下：

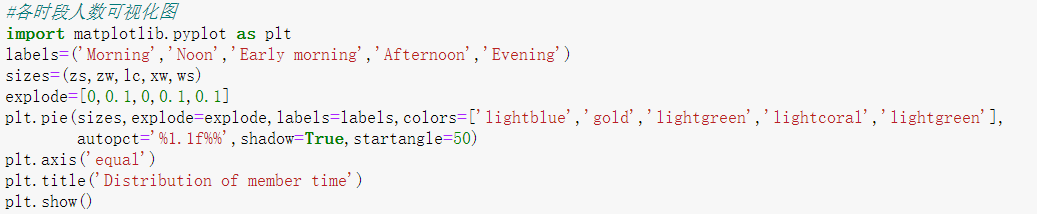
首先了解每个时间段的具体产生订单数量：



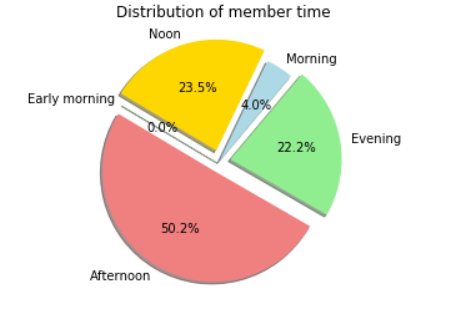
运行结果如下：



为了直观了解各时间段的占比，绘制饼图如下：



运行结果如下：

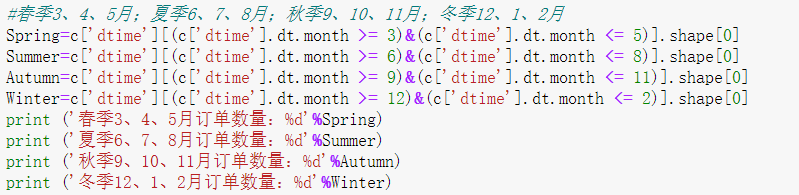


通过上面的可视化结果，可以发现，凌晨的顾客几乎可以当作没有太多顾客，只有不到100单的订单，而在下午，订单数量占了一整天的一半，因此，在制定工作人员时间和策略上，可以在凌晨缩减工作人员数量，把人力集中安排在中午、下午和晚上，尤其是下午。

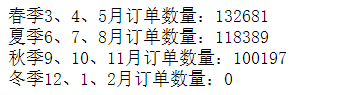
**4.2.10 各季节会员消费人数**

因为对于商场而言，进货的时间和频率通常和季节有关，不同季节的进货数量的设置不同影响着商场的运营成本，为此，特别统计不同季节的销量，过程如下：

首先了解每个季节的具体产生订单数量：



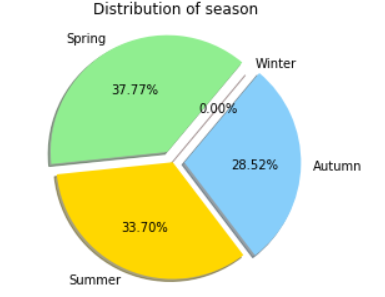
运行结果如下：



为了直观了解各季节销量的占比，绘制饼图如下：



运行结果如下：



通过观察上图可以发现，冬季的销量几乎是冰点，可能因为正处过年时节，天气寒冷，员工需要回家，客户也不在此时购物，作为商家，可以尽量避免在冬季囤积大量的商品，尽量在入冬前售出，又因为春季的订单最多，所以在快入春时可以提前入货。

## 总结

结合可视化的分析结果，我们初步为商场提供以下策略建议：

**·**该商场的会员大多数是青年人和中年人，占到95.4%，为此，商场在制定策略时，可以有倾向地把主要客户定位在青年人和中年人；

**·**该商场的89.6%主要顾客是女性顾客，因而在考虑商场商品的种类时，应当尽可能多地出售女性商品从而获得更高的利润；

**·**尽管青年人和中年人会员占比相近，但是中年人的消费金额占比比青年人多，占比有53.2%，建议出售更多中年人的产品，吸引中年人；

**·**相比不同性别会员数量的占比，消费金额占比女性稍微下降，原来为89.6%，如今为87.6%，可见，男性的消费金额就个体而言比女性要高，但总体达不到女性水平，可以发展更多的男性客户；

**·**消费金额来说，会员比非会员消费金额的总数要大，尽管非会员产生的订单消费记录比较多，因此，就商场的长远规划来说，应当尽可能发展多的会员；

·除了2018年因为只有1个月的数据所以数值较小，该商场的会员每年的消费金额都在上升，可见发展会员是有助于商场长期的发展的，所以可以尽可能发展会员；

·除了2015年的数据较为反常，每年的2月份几乎都是每年中的历史最低点，因为正处过年时节，此时1月份会相对较高，因为大家需要购置年货，作为商家，可以尽量避免在2月份引进大量的商品，在1月或年末11-12月引进商品，避免导致滞销和过期；

·凌晨的顾客几乎可以当作没有太多顾客，只有不到100单的订单，而在下午，订单数量占了一整天的一半，因此，在制定工作人员时间和策略上，可以在凌晨缩减工作人员数量，把人力集中安排在中午、下午和晚上，尤其是下午；

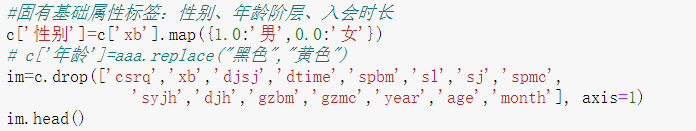
·冬季的销量几乎是冰点，可能因为正处过年时节，天气寒冷，员工需要回家，客户也不在此时购物，作为商家，可以尽量避免在冬季囤积大量的商品，尽量在入冬前售出，又因为春季的订单最多，所以在快入春时可以提前入货。

# 用户画像

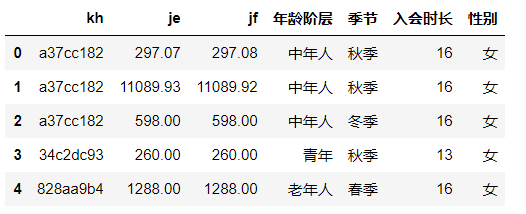
为了能够绘制用户画像，我们建立了三类特征标签，分别是用户基本特征标签、用户业务特征标签、用户偏好特征标签。最终构建一个只含有以上三类标签的表hy。

## 构建会员基本特征标签

根据用户固有基础属性标签，我们构建会员基本特征标签有：性别、年龄、入会时长；过程如下：



基本处理后前五数据如下：

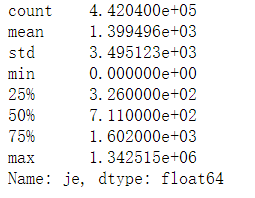


## 构建会员业务特征标签

根据用户属性标签，我们构建会员业务特征标签有：消费水平、新老会员；过程如下：

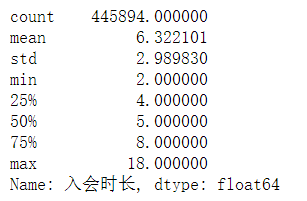
以下所有标签，对于三分的，采用采用区位25%、50%、75%作为分界点；对于二分的，采用50%作为分界点。

为了确定消费水平低、中等、高的区间，采用区位25%、50%、75%作为分界点，过程如下：



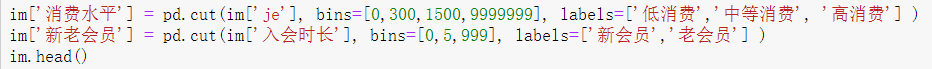
根据描述，确定300以下为低消费；300-1500为中等消费；1500以上为高消费；

为了确定入会时长构成新会员还是老会员，采用50%作为分界点，过程如下：



从而确定5年以下为新会员，5年以上为老会员；

数据处理过程如下：



运行结果如下：



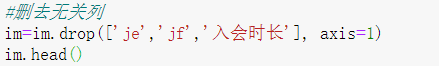
## 构建会员偏好特征标签

根据用户属性标签，我们构建会员偏好特征标签有：购物时间段、偏好季节、商品；过程如下：

偏好季节对应季节变量，商品也对应spmc变量，不进行处理，在这里增加购物时间段列，过程如下：



运行后，结果前五列如下：

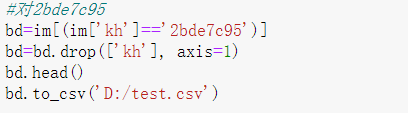


删去无关列后结果如下：



## 建立用户画像

以2bde7c95用户为例，建立其用户画像，首先将其数据保存到csv中：



使用wordcloud建立词云：



运行后如图所示：



可见用户2bde7c95是老会员，而且是青年，习惯中午进行消费，而且是低消费类型。

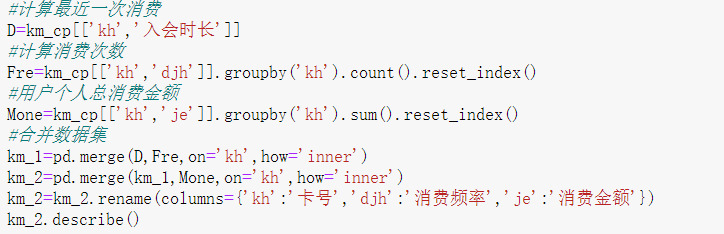
# 用户精细化分和K-means聚类

## 基于LFM模型对用户进行化分

**6.1.1 LFM模型的构建**

根据RFM模型，分析会员最近一次消费时间、消费频率以及消费金额，把三个指标作为衡量会员价值的依据，最近一次消费时间与会员价值成反比，（一定周期）消费频率、（一定周期）消费金额与会员价值成正比。而由于数据的特殊，基于RFM模型改进得到DFM模型，最近一次消费时间变成入会时间进行分析。

通过收集相关资料，还可以把最大单笔消费金额、特价商品消费占比、高单价商品消费占比作为另外三个衡量指标，这里暂不涉及。



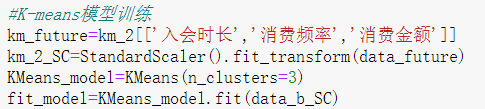
运行结果如下：



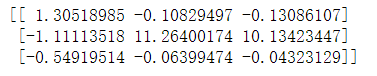
## 基于K-means算法对用户进行聚类

k-means聚类是常用的聚类分析算法，聚类效果达到中上水平，k-means是一种基于原型的、根据距离划分组的算法，其时间复杂度比其他聚类算法低，在这里，我指定划分组的个数k=3。

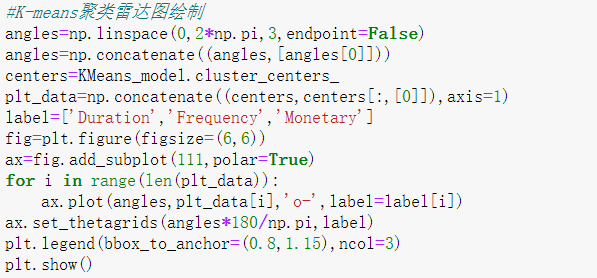
构建K-Means聚类模型：



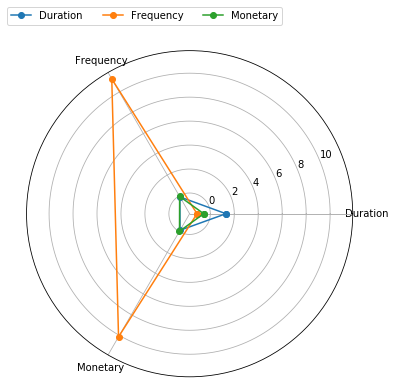
输出模型聚类中心结果如下：



基于K-means聚类绘制雷达图：



绘制结果如下：



## 小结

结合聚类中心和雷达图的分析结果，我们初步为商场提供以下策略建议：

**·**就Duration而言，对活跃的客户采取密集的营销信息推送；对沉默客户减少推送频率，提升优惠力度；对睡眠客户，大型活动时进行营销推送；对于流失客户，在超大型活动中（如双十一）进行推送；

**·**就Frequency而言，对新客户传递促销信息；对老客户传递品牌信息；对成熟客户传递新品/活动信息；对忠实客户传递会员活动和权益信息；

·就Monetary而言，对低贡献客户和中贡献客户，提供促销商品/折扣活动；对中高贡献客户，提供形象商品/品牌活动。