超市销售数据分析

近年来，随着新零售业的快速发展，消费者购买商品时有了更多的对比和选择，导致超市行业的竞争日益激烈，利润空间不断压缩。超市的经营管理产生了大量数据，对这些数据进行分析，可以提升超市的竞争力，为超市的运营及经营策略调整提供重要依据。

**学习目标**

1. 掌握pandas对数据预处理。
2. 掌握pandas统计数据并且分类。
3. 掌握matplotlib.pyplot的基本画图方法。

# 读取超市零销售数据并做预处理

**任务描述**

首先利用pands库读取我们零销售数据，pandas库是python中比较友好的做数据预处理库，可以读取mysql，csv，txt，json等类型的数据。本次实验的数据是存放在一个csv中。我们读取到数据后，需要对数据来进行初步的了解，发现数据是一个42816行，17列的一个数据，其中列中用顾客编号，食品编号，销售时间，商品单价，是否促销等属性。而有些属性其实是比较冗余的，如小类编码和小类名称其实要其中一个就行。因为存在这一类的问题，我们不能直接采取现有的数据进行操作，而是需要进行预处理。

**任务分析**

可以通过以下2个步骤了解对商场零售数据的预处理。

1. 了解商场零售数据数据。
2. 删除一些冗余的数据属性，格式化日期等等。

## 了解商场零售数据数据

本案例主要对商场零售数据进行预处理，统计，分析。部分数据如表 1‑1所示。

表 1‑1 商品零售数据样式

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **顾客编号** | **大类编码** | **大类名称** | **中类编码** | **中类名称** | **小类编码** | **小类名称** | **销售日期** | **销售月份** | **商品编码** | **规格型号** | **商品类型** | **单位** | **销售数量** | **销售金额** | **商品单价** | **是否促销** |
| 0 | 12 | 蔬果 | 1201 | 蔬菜 | 120109 | 其它蔬菜 | 20150101 | 201501 | DW-1201090311 |  | 生鲜 | 个 | 8.0 | 4.0 | 2.0 | 否 |
| 1 | 20 | 粮油 | 2014 | 酱菜类 | 201401 | 榨菜 | 20150101 | 201501 | DW-2014010019 | 60g | 一般商品 | 袋 | 6.0 | 3.0 | 0.5 | 否 |
| 2 | 15 | 日配 | 1505 | 冷藏乳品 | 150502 | 冷藏加味酸乳 | 20150101 | 201501 | DW-1505020011 | 150g | 一般商品 | 袋 | 1.0 | 2.4 | 2.4 | 否 |

## 删除一些冗余的数据属性，格式化日期

删除一些冗余的数据属性，格式化日期，具体流程如下。

1. 删除一些冗余的列，如：'大类名称','中类名称','小类名称','单位','销售月份'。
2. 格式化日期。
3. 将数据保存为task1\_1.csv。

# 统计其他列对销售金额的影响

**任务描述**

该任务需要对数据比较了解，熟悉并掌握pandas的分类方法，我们需要统计大类商品，是否促销，商品的类别等因素对销售有多少影响。

**任务分析**

统计各因素下销售金额的具体流程如下。

1. 统计每个大类商品的销售金额，将结果保存为“task1\_2.csv。
2. 统计每个中类商品的促销销售金额和非促销销售金额，将结果保存为“task1\_3.csv。
3. 统计生鲜类产品和一般产品的每周销售金额，将结果保存为“task1\_4.csv。
4. 统计每位顾客每月的消费额及消费天数，将结果保存为“task1\_5.csv。

## 统计每个大类商品的销售金额

使用pandas对数据进行分类统计，首先按商品大类来进行统计，大类一般包括‘蔬果’，‘粮油’，‘烘焙’等类型。对此分析可以让我们知道那一大类比较受欢迎，在此由于大类名称和大类编码表示的意思一致，由此我们保留了其中之一的大类编码，用一下代码进行统计：data['大类编码销售金额']=data.groupby('大类编码').agg({'销售金额':'sum'})，得到的结果如图2.1所示。

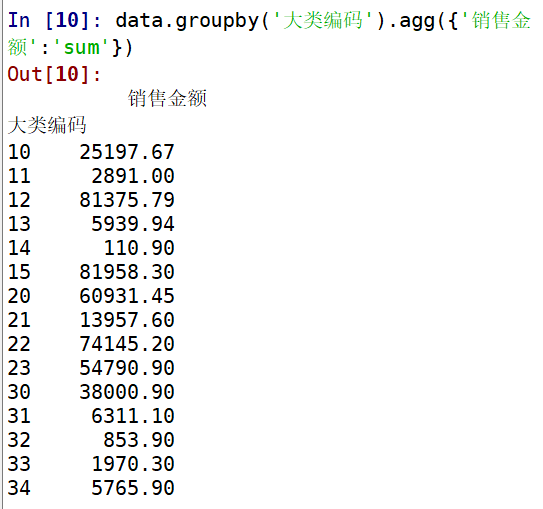


图 2‑1 大类编码统计销售额

由图我们可以看出编码为‘15’，‘20’，‘22’的大类比较受欢迎，其对应的大类名称为‘日配’，‘粮油’，‘休闲’。相反，不那么受欢迎的编码是‘14’，‘32’，其对应的名称为‘烘焙’，‘家电’。因此我们建议商场对‘日配’，‘粮油’，‘休闲’的大类商品进行入货。

## 统计每个中类商品的促销销售金额和非促销销售金额

我们统计完大类商品后，当然也会按中类商品进行统计。而且按照销售和非销售金额进行统计。中类名称是‘冷藏乳品’，‘蔬菜’，‘卫生巾’等分类。按照此分类可以更加细致的看出哪些商品更受欢迎。并且按照是否销售进行统计可以看出销售与非销售间的区别。因此我给出的代码是：d = data.groupby(['是否促销','中类编码']).agg({'销售金额':'sum'})。我们统计到的部分数据如下图2.2

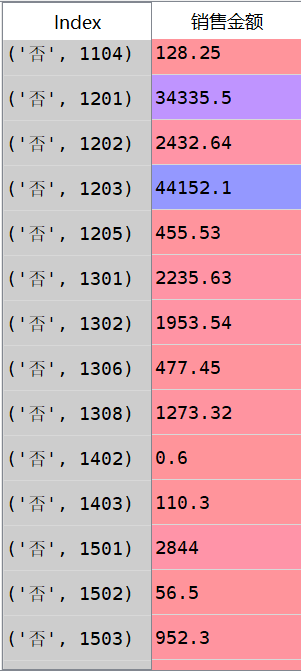


图 2‑1 中类编码是否销售统计销售额

## 统计生鲜类产品和一般产品的每周销售金额

生鲜类和一般类的产品是我们的商场最常看到的产品，因此我们继而对他们进行统计，而且周期为每周，而不是每个月，也不是每天，这样周期不长不短，能更好的说明销售的情况。由于我们的数据框里没有每周属性，但有日期的属性，所以我们可以将日期格式化成pd.to\_datetime类型的日期，以方便我们处理数据，datetime类型的数据可以很轻易地获取到当天是一年中的那一周。

首先我们需要对原来是objet的日期类进行格式化。在此之前，我们发现有的日期错乱了，不符合真实的日期范围，因此我们需要将其找出来并删掉。代码如下：

|  |
| --- |
| #获取有异常的日期  excriqi = set()  for i in data['销售日期']:  try:  pd.to\_datetime(i)    except:  print(i)  excriqi.add(i) |

由此我们捕获到了异常日期2015-2-29。删除异常日期，并且格式化日期，代码如下：

|  |
| --- |
| #格式化日期  def formatTime(x):  x=str(x)  return str(x[:4])+'-'+str(x[4:6])+'-'+str(x[6:])  data['销售日期']=data['销售日期'].apply(formatTime)  #去掉有异常的日期  for i in excriqi:  data.drop(data[data['销售日期']==i].index,inplace=True)  #将日期的格式转换好  data['销售日期']= pd.to\_datetime(data['销售日期']) |

格式化完成后，我们可以很轻易地获取每天在一年中是哪个月了，代码如下：

|  |
| --- |
| data['第几周']= pd.DatetimeIndex(data['销售日期']).weekofyear |

进而统计中类商品每周销售金额，代码如下：

|  |
| --- |
| d = data.groupby(['商品类型','第几周']).agg({'销售金额':'sum'}) |

由此我们也得到了统计的数值，部分数据如下图2.3所示。其中index即为商品类型和第几周

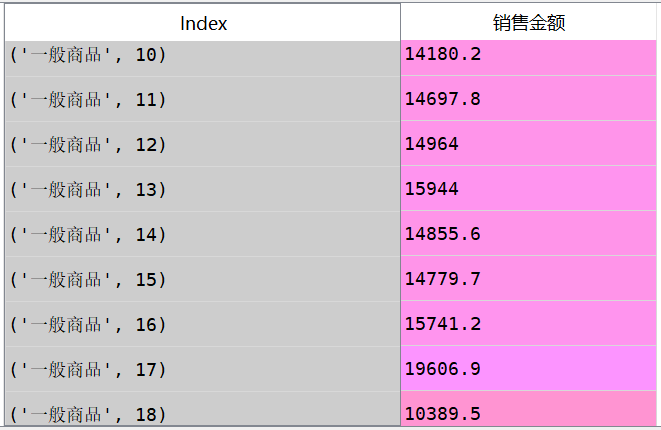


图 2‑3 生鲜类产品和一般产品的每周销售金额

## 统计每位顾客每月的消费额及消费天数

除了对商品方面进行统计之外，我们还需要对顾客本身进行统计，查看顾客每个月的消费额，及其消费天数。由此可以更清晰知道那些顾客经常光顾商店。对此我给出的代码如下：

|  |
| --- |
| d = data.groupby(['顾客编号']).agg({'销售金额':'sum','销售日期':'count'}) |

统计得到的部分数据如下图2.4，其中index即顾客编号和月份。

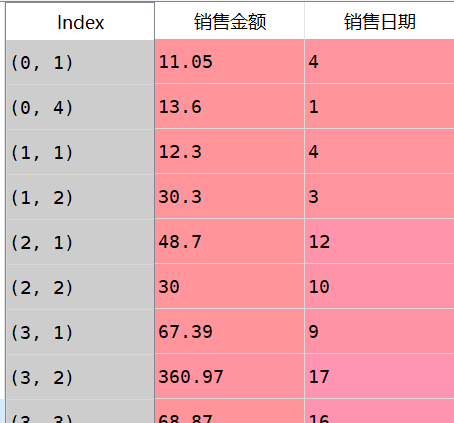


图 2‑3 每位顾客每月的消费额及消费天数

# 绘制商品销售金额情况图，分析销售状况

**任务描述**

对于前面的一些任务，我们仅仅只是得到了一些统计的数据，我们没有直观的看到数据的走向，以及数据描述商品销售的情况，因此我们需要用到matplotlib.pyplot下的画图工具将我们分析统计得到的数据画成折线图，散点图，饼状图等等，由此更加直观的知道那些数据比较显著表达问题。

**任务分析**

主要对以下几个数据进行绘图。

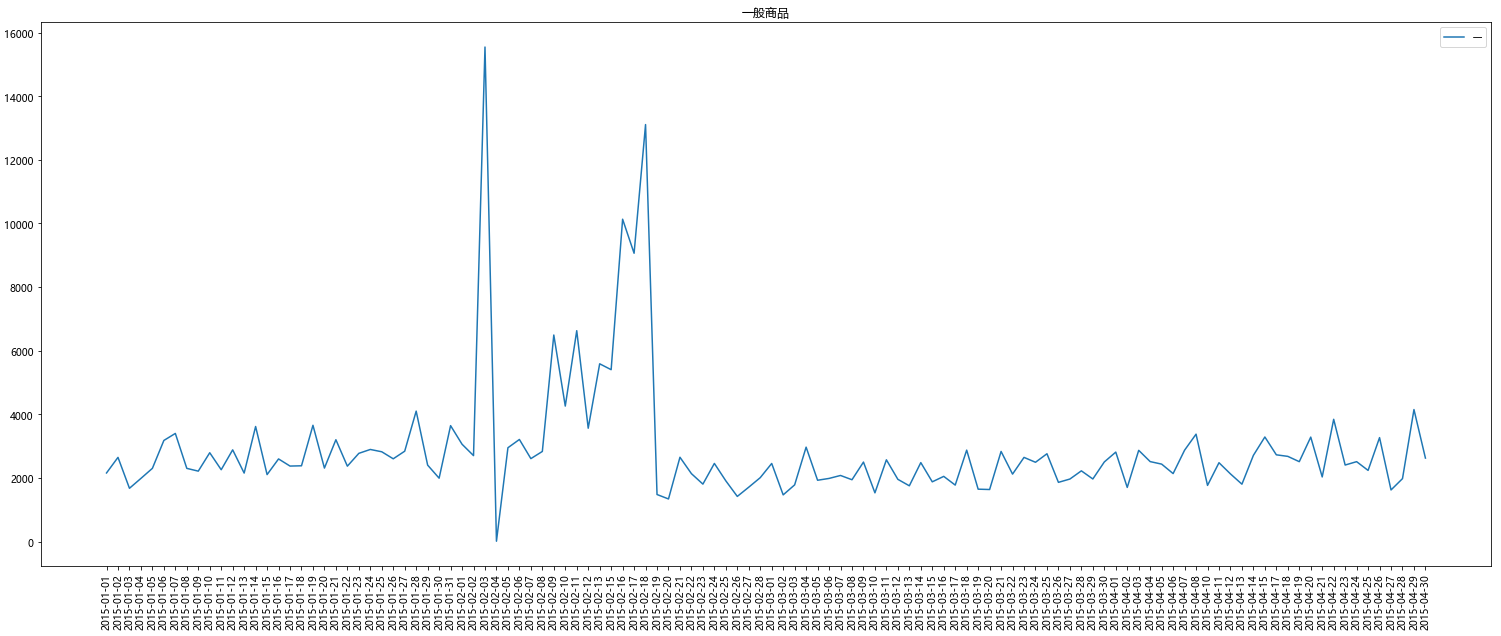
1. 绘制生鲜类商品和一般商品每天销售金额的折线图，并分析比较两类产品的销售状况。
2. 按月绘制各大类商品销售金额的占比饼图，并分析其销售状况。
3. 绘制促销商品和非促销商品销售金额的周环比增长率柱状图。
4. 根据消费情况，分别为累计消费前 10 的顾客画像

## 绘制生鲜类商品和一般商品每天销售金额的折线图

首先我们一样要读取数据，然后将数据商品类型分类统计其每天销售金额。然后获取生鲜类商品和一般商品的每天销售金额，因为两种销售金额的天数不一致，因此这里我们需要两个子图描绘两种商品的销售金额状况。具体代码如下：

|  |
| --- |
| data = pd.read\_csv('../01.中间数据（如有）/task1\_2.csv')  d = data.groupby(['商品类型','销售日期'],as\_index=False).agg({'销售金额':'sum'})  ax1 = d.loc[d['商品类型']=='一般商品',:]  ax2 = d.loc[d['商品类型']=='生鲜',:]  import matplotlib.pyplot as plt  # 一般商品每天销售金额的折线图  plt.figure(figsize=(26,10))  plt.title(list(ax1['商品类型'])[0])  plt.plot(ax1['销售日期'],ax1['销售金额'])  plt.xticks(rotation=90)  plt.legend(list(ax1['商品类型'])[0])  # 生鲜类商品每天销售金额的折线图  plt.figure(figsize=(26,10))  plt.title(list(ax2['商品类型'])[0])  plt.plot(ax2['销售日期'],ax2['销售金额'])  plt.xticks(rotation=90)  plt.legend(list(ax2['商品类型'])[0]) |

由以上代码得到我们的图像如下图3.1。



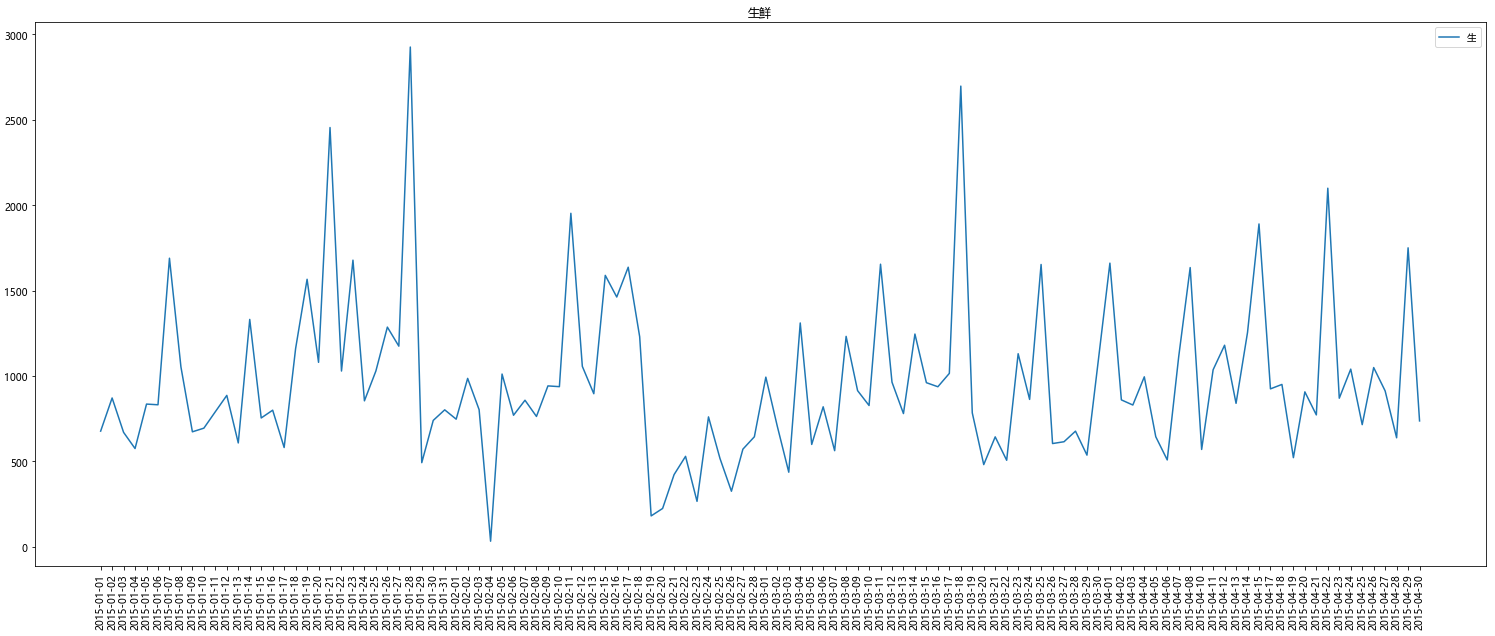


图 2‑3 生鲜类商品和一般商品每天销售金额的折线图

## 按月绘制各大类商品销售金额的占比饼图

按照每个月查看个大类商品的销售金额的比值，可以更直观的比对出哪一大类的商品销售的更受欢迎。首先一样需要读取数据，同时给数据加上月份的属性，然后再对商品类型，月份进行组合，对销售金额进行统计。由此我们给出的代码如下：

|  |
| --- |
| # In[任务3.2 按月绘制各大类商品销售金额的占比饼图，并分析其销售状况]  import matplotlib.pyplot as plt  data = pd.read\_csv('../01.中间数据（如有）/task1\_2.csv')  data['monthMany'] = pd.DatetimeIndex(data['销售日期']).month  d = data.groupby(['monthMany','大类编码'],as\_index=False).agg({'销售金额':'sum'})  fig = plt.figure(figsize=(16,16)) #创建画布  ax1 = fig.add\_subplot(2,2,1)  plt.title('第一季度')  ax1.pie(d.loc[d['monthMany']==1,'销售金额'],autopct='%.2f %%',labels=d.loc[d['monthMany']==1,'大类编码'])  ax2 = fig.add\_subplot(2,2,2)  plt.title('第二季度')  ax2.pie(d.loc[d['monthMany']==2,'销售金额'],autopct='%.2f %%',labels=d.loc[d['monthMany']==2,'大类编码'])  ax3 = fig.add\_subplot(2,2,3)  plt.title('第三季度')  ax3.pie(d.loc[d['monthMany']==3,'销售金额'],autopct='%.2f %%',labels=d.loc[d['monthMany']==3,'大类编码'])  ax4 = fig.add\_subplot(2,2,4)  plt.title('第四季度')  ax4.pie(d.loc[d['monthMany']==4,'销售金额'],autopct='%.2f %%',labels=d.loc[d['monthMany']==4,'大类编码']) |

由以上代码得到的图像如图3.2所示。

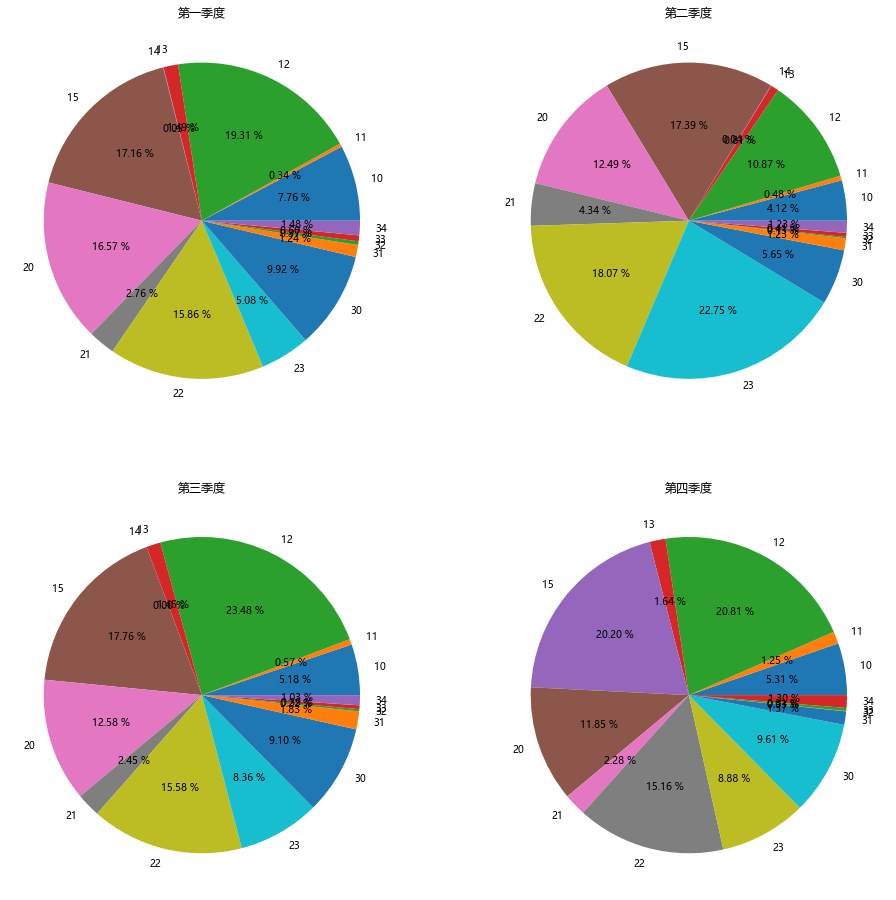


图 3‑2 每月各大类商品销售金额的占比饼图

由图可以清楚的看出几乎每个季度，大类编码为‘12’，‘20’的商品销售额占比都比较高。大类编码为‘21’的商品在第四个季度里，销售额远不如前三个季度。编号为‘23’的销售额仅在第二季度比较大。

## 绘制促销商品和非促销商品销售金额的周环比增长率柱状图

对比完季节性之间的差异后，我们还没对促销与非促销的数据进行统计分析，所以我们需要对促销和非促销进行分析，同时需要看出其两者间的发展趋势，因此我们需要算出促销和非促销的周环比，我们知道周环比公式为：周环比增长率=（本周数-上周数）/上周数×100%。因此我们首先需要有周的属性，同样需要通过日期得到周，然后对周和是否促销进行分组，对销售金额进行统计。

统计完金额后就需要按周来计算周环比了，下面我给出的代码：

|  |
| --- |
| #读取数据  data = pd.read\_csv('../01.中间数据（如有）/task1\_2.csv')  data['第几周']= pd.DatetimeIndex(data['销售日期']).weekofyear  d = data.groupby(['是否促销','第几周'],as\_index=False).agg({'销售金额':'sum'})  #周环比增长率=（本周数-上周数）/上周数×100%  #由于第一周的数据没有，因此不需要算第一周  def zhouhuanbi(x):  a1 = x['销售金额'][1:].reset\_index().drop('index',axis=1)  a2 = x['销售金额'][:-1].reset\_index().drop('index',axis=1)  o = (a1 - a2) / a2  # print(o['销售金额'])  o = list(o['销售金额'])  o.insert(0,None)  return o  yes = d.loc[d['是否促销']=='是',:]  no = d.loc[d['是否促销']=='否',:]  y = zhouhuanbi(yes)  n = zhouhuanbi(no)  d['周环比'] = [None,None] + y + n  dd = d[['是否促销','第几周','周环比']].dropna() |

得到的dd即是我们需要的数据，我们将数据进行绘图，更直观的看到促销与非促销之间的发展趋势。代码如下：

|  |
| --- |
| #绘图  fig = plt.figure(figsize=(10,8)) #创建画布  plt.bar(list(dd.loc[dd['是否促销']=='否','第几周']-0.2),list(dd.loc[dd['是否促销']=='否','周环比']),alpha=0.6,width=0.4)  plt.bar(list(dd.loc[dd['是否促销']=='是','第几周']+0.2),dd.loc[dd['是否促销']=='是','周环比'],width=0.4)  plt.xticks(list(dd.loc[dd['是否促销']=='是','第几周']))  plt.xlabel('周')  plt.ylabel('周环比')  plt.legend(['非促销','促销']) |

由代码可得图3.3，由图可知非促销的周环比更大，因此我们给超市的建议是尽量少搞促销。

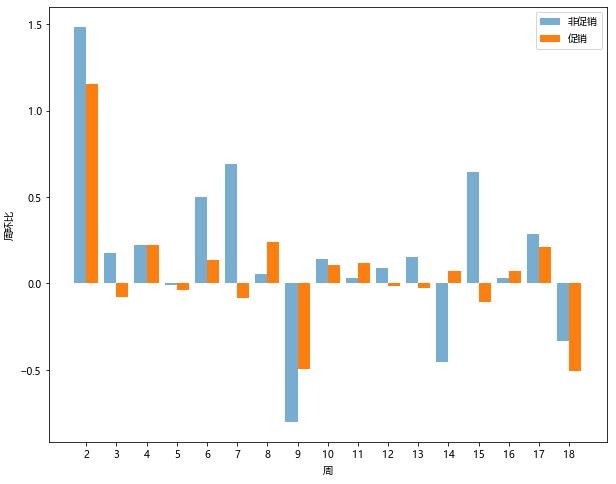


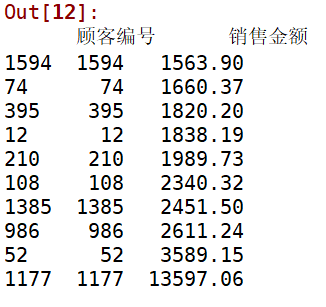
图 3‑3 促销商品和非促销商品销售金额的周环比增长率柱状图

## 根据消费情况，分别为累计消费前 10 的顾客画像

为统计消费前十的顾客画出图像我们可以看出在此消费最大或将近比较大的数目是多少。首先我们需要对顾客编号进行分组，然后对销售金额进行统计。最后再按照销售金额进行排序得到最终想要的结果。代码如下：

|  |
| --- |
| # In[课时 11 : 任务4.1 根据消费情况，分别为累计消费前 10 的顾客画像]  import matplotlib.pyplot as plt  import pandas as pd  #读取数据  data = pd.read\_csv('../01.中间数据（如有）/task1\_2.csv')  #统计顾客消费金额  d = data.groupby(['顾客编号'],as\_index=False).agg({'销售金额':'sum'})  #排序  d.sort\_values('销售金额', inplace=True)  #找到前十的顾客  qianshi = d.iloc[-10:,:]  #画图象  fig = plt.figure(figsize=(10,8)) #创建画布  plt.plot(range(len(qianshi)),qianshi['销售金额'],marker='o')  plt.xticks(range(len(qianshi)),qianshi['顾客编号'])  plt.xlabel('顾客编号')  plt.ylabel('销售金额') |

可以看到我们得到的前十名销售金额数据如下：



画出的图像如下图3.4

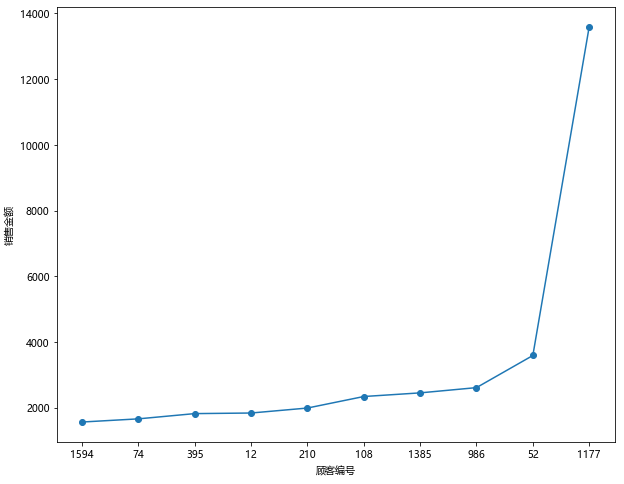


图 3‑4 累计消费前 10 的顾客

# 小结

本案例分析统计了商场商品的销售情况，在大类，中类，小类以及是否促销，按时间方面都有进行分析与统计，很好的分析说明了商场的销售状况。