**基于Python + Pandas + Matplotlib + Seaborn的**

**数据分析**

[一、引言 3](#_Toc148269230)

[二、原始数据 3](#_Toc148269231)

[三、数据处理 4](#_Toc148269232)

[四、数据可视化 8](#_Toc148269233)

[五、数据总结 12](#_Toc148269234)

[六、学习总结 12](#_Toc148269235)

一、引言

本年数学建模竞赛刚结束不久，为学习数据处理的使用，故使用C题的部分数据（附件1、附件2、附件3）进行数据处理及分析以达到学习与实践。

二、原始数据

对于数据的读取，由于附件为Excel表格，故采用Pandas来进行读取，示例代码如下：



以下分别为附件1、附件2、附件3的部分内容：

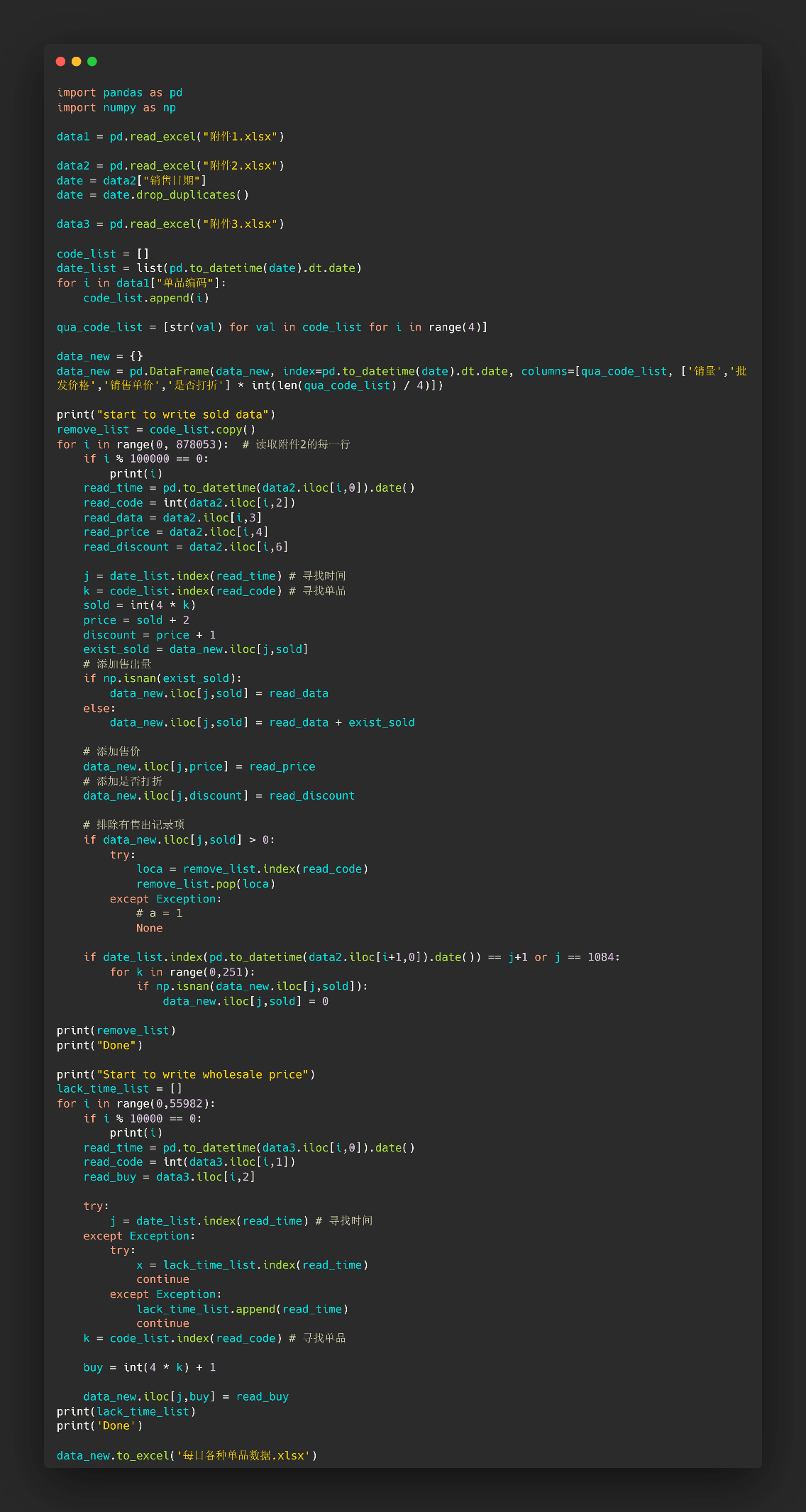




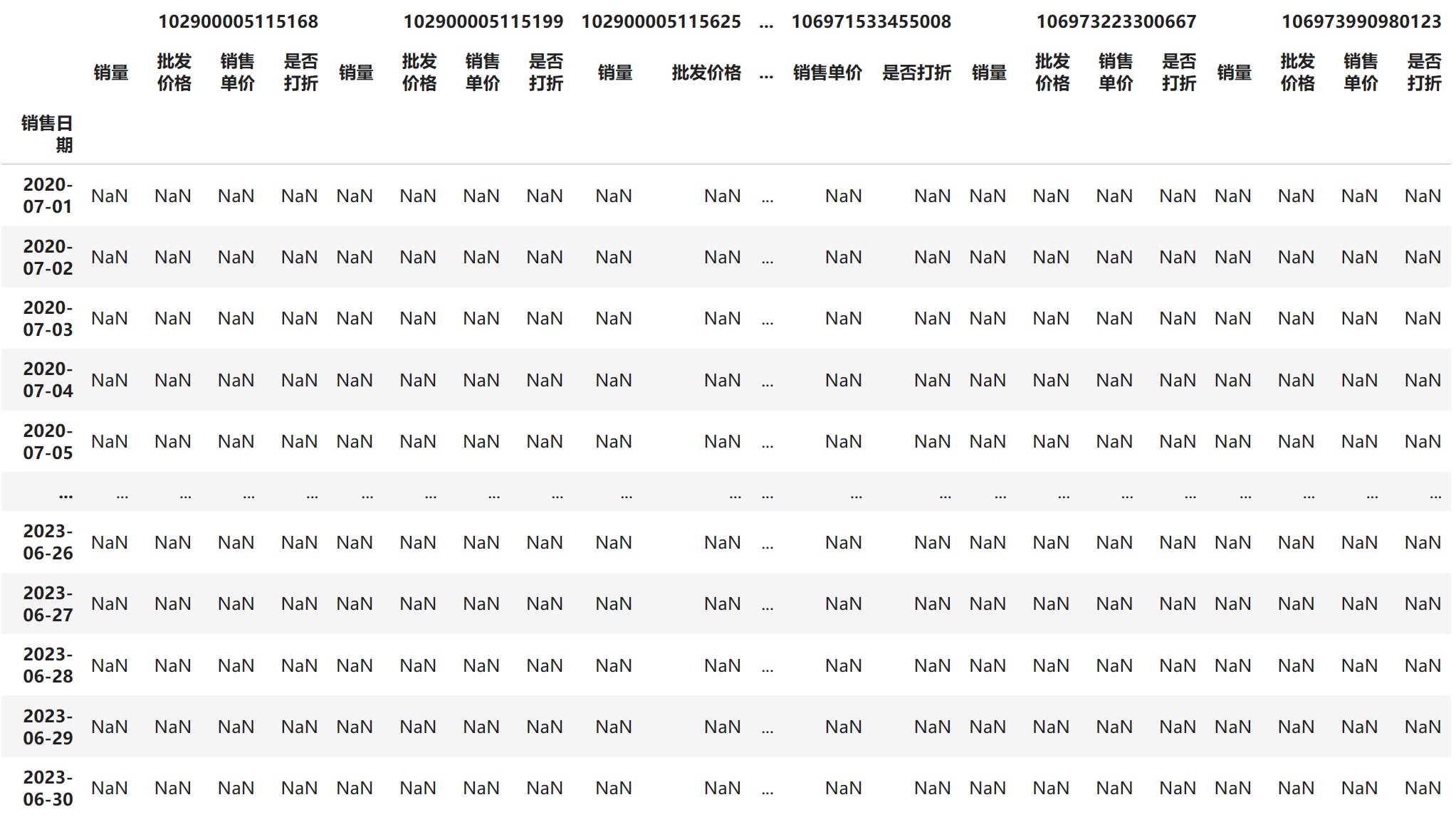


三、数据处理

首先对附件2进行处理：将其中销量的数据按照一天来进行加和，将每一种，每一天的销量以及销售单价提取出来；然后对附件3进行处理：提取出每一天各商品的批发价格。将以上的数据进行整合，列的索引为商品编码，顺序根据附件1，行的索引为日期（即某年某月某日），内容参照附件2的日期。整合后将数据以Excel形式进行保存。代码如下：



得到的数据部分如下:

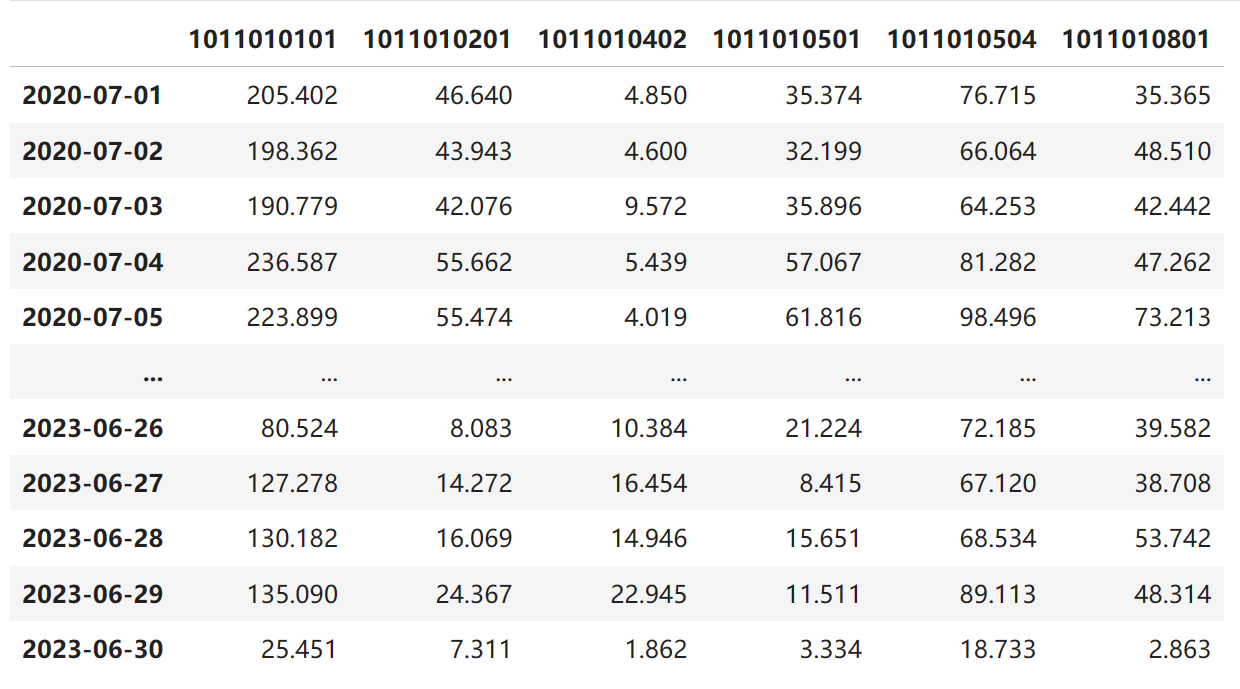


该部分内容为NaN，说明在该段时间该种物品没有进行销售和批发购买。

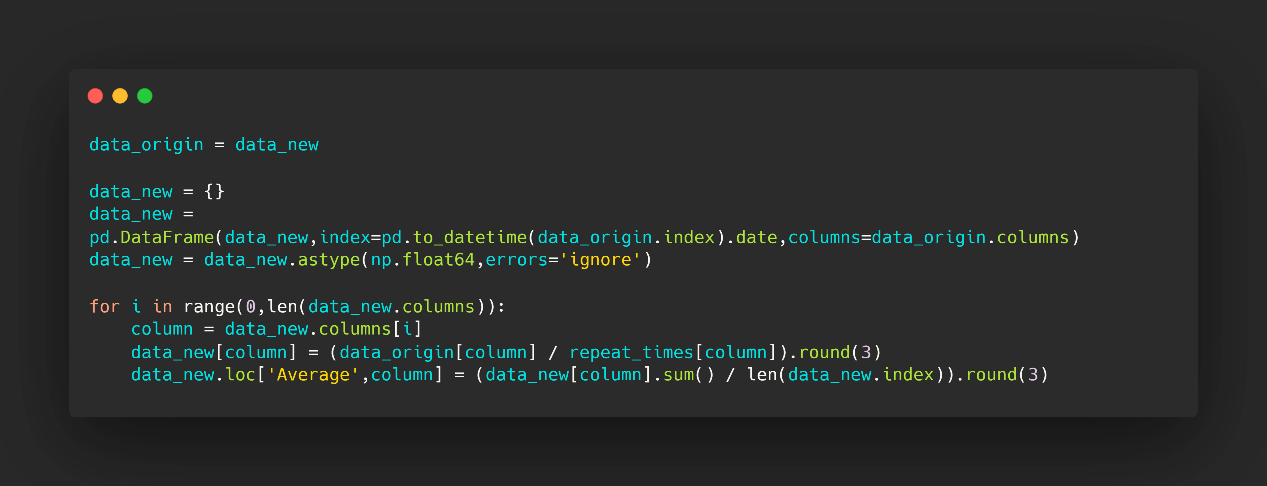
针对以上数据，将销量再次提取进行处理：按照附件1中的品类进行加和，以观察不同品类的销量，代码如下：



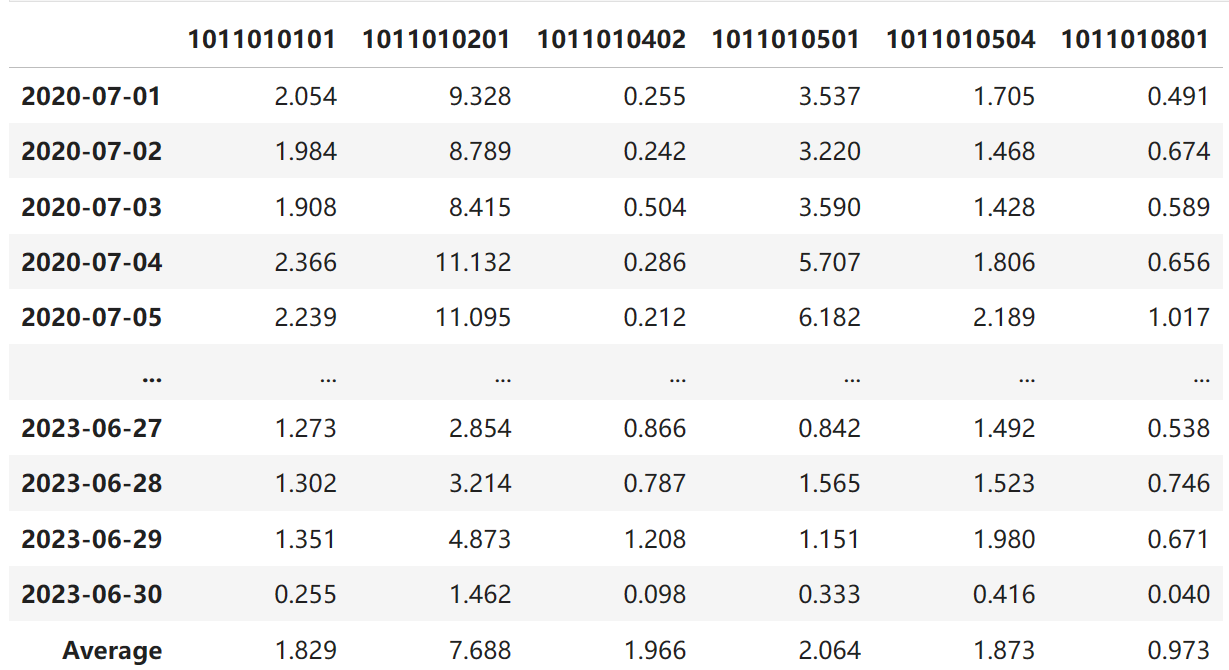
得到的部分数据如下：



由于每个品类所含的商品数量不同，故销量无法用作比较，因此，对其数据进行了求平均（首先按照“每一天各品类的销量/各品类所含商品的个数”进行计算，然后使用该数据进行“每天平均各品类的销量/统计日期的天数”计算），代码承接上一步的代码，展示如下：



得到的部分数据如下：



处理结束后，将以上两数据以“Sheet1”“Sheet2”的命名保存至同一Excel文件中，命名为“每日各种类型出售量”，代码如下：



至此，数据处理结束。

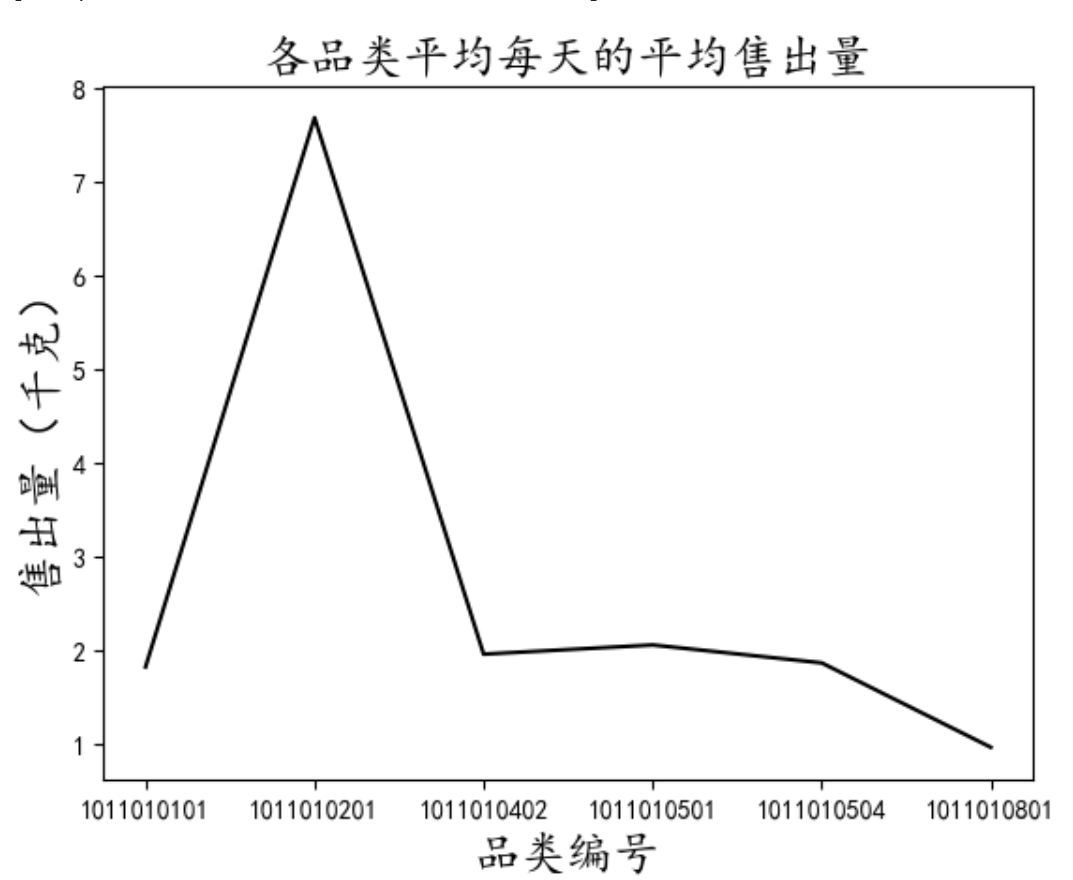
四、数据可视化

通过数据处理后，分别得到了各商品每天的销量、销售单价、批发价格，各品类每天的平均销量、平均至每天的销量，接下来利用Matplotlib和Seaborn进行数据可视化以得到直观的结果。

首先对各品类平均每天的平均销量进行数据可视化，代码如下：



得到的图像如下：

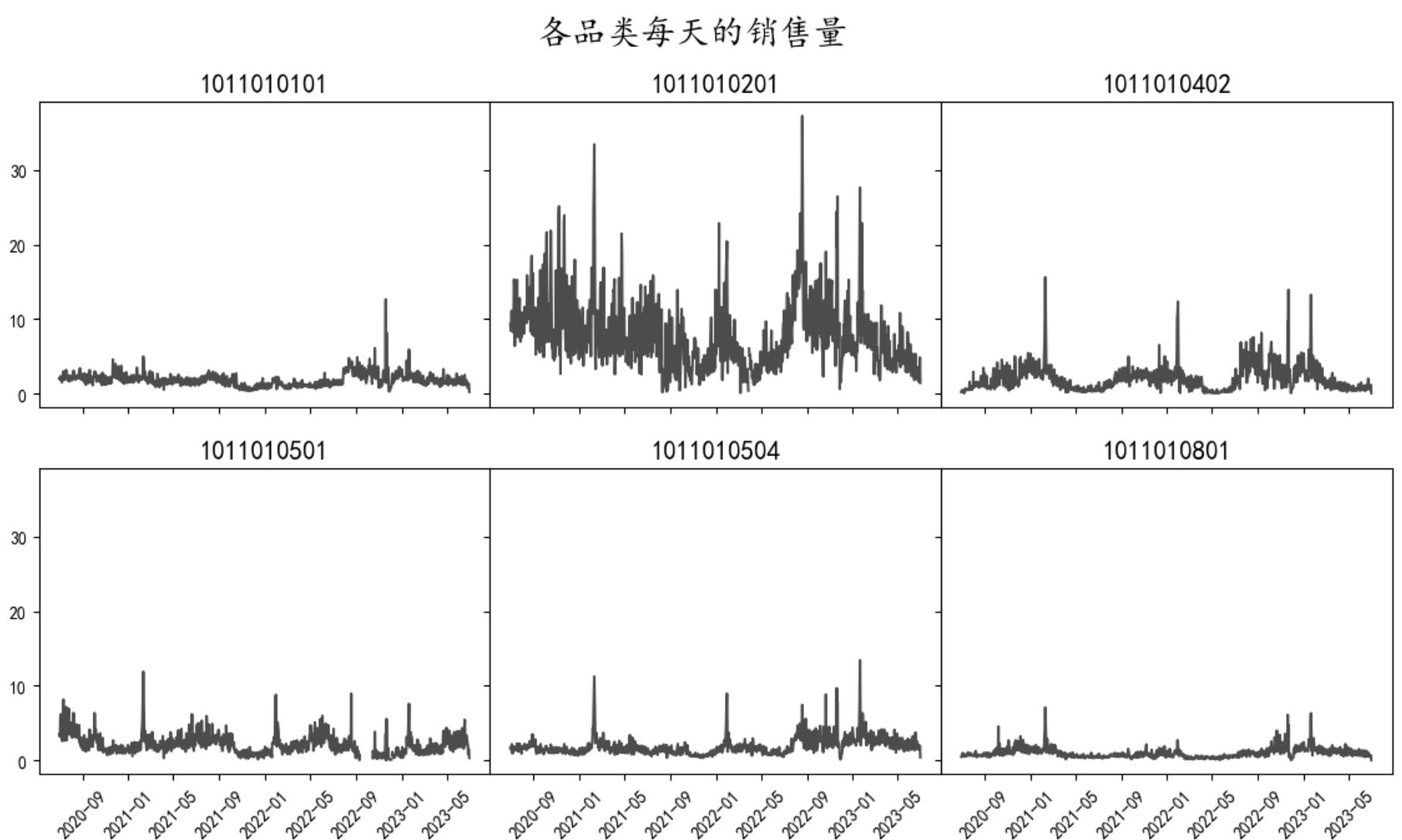


由此可观察到，品类编号为1011010201的商品的销量最高，经在附件1中的查询，该品类为花菜类。

接下来对各品类的平均销量进行数据可视化，数据接前一部分代码如下：



由此得到的图像如下：

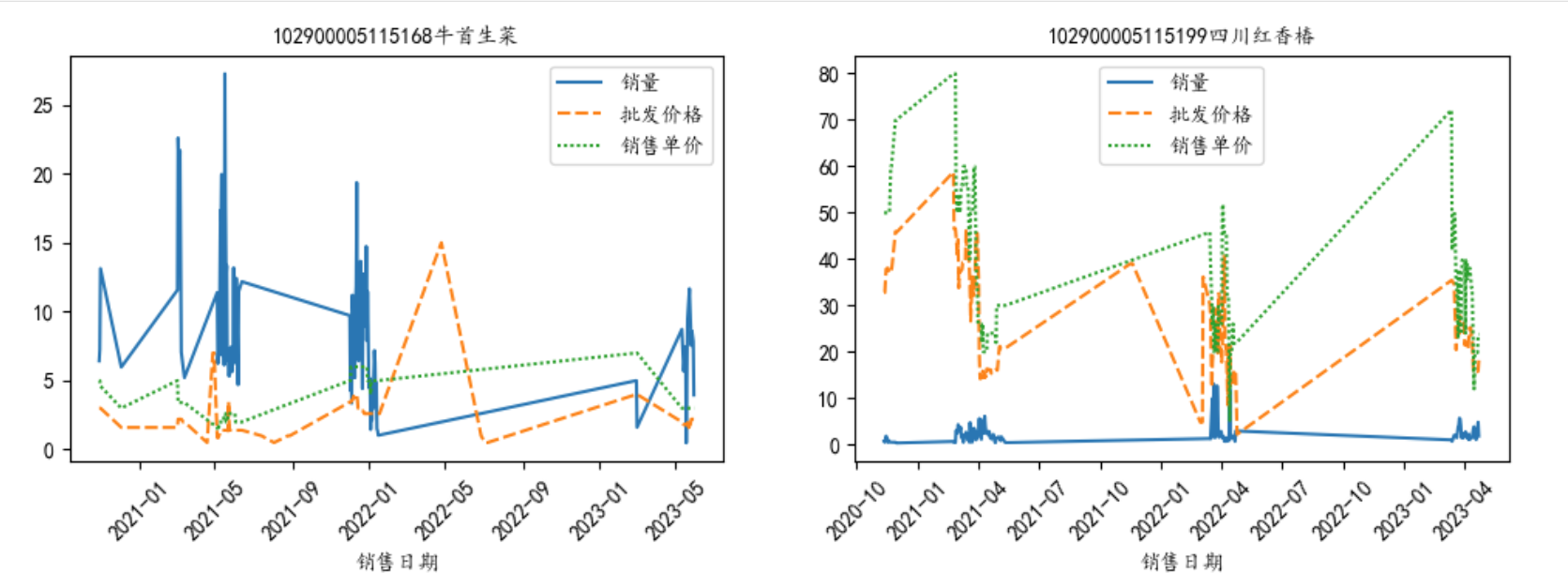


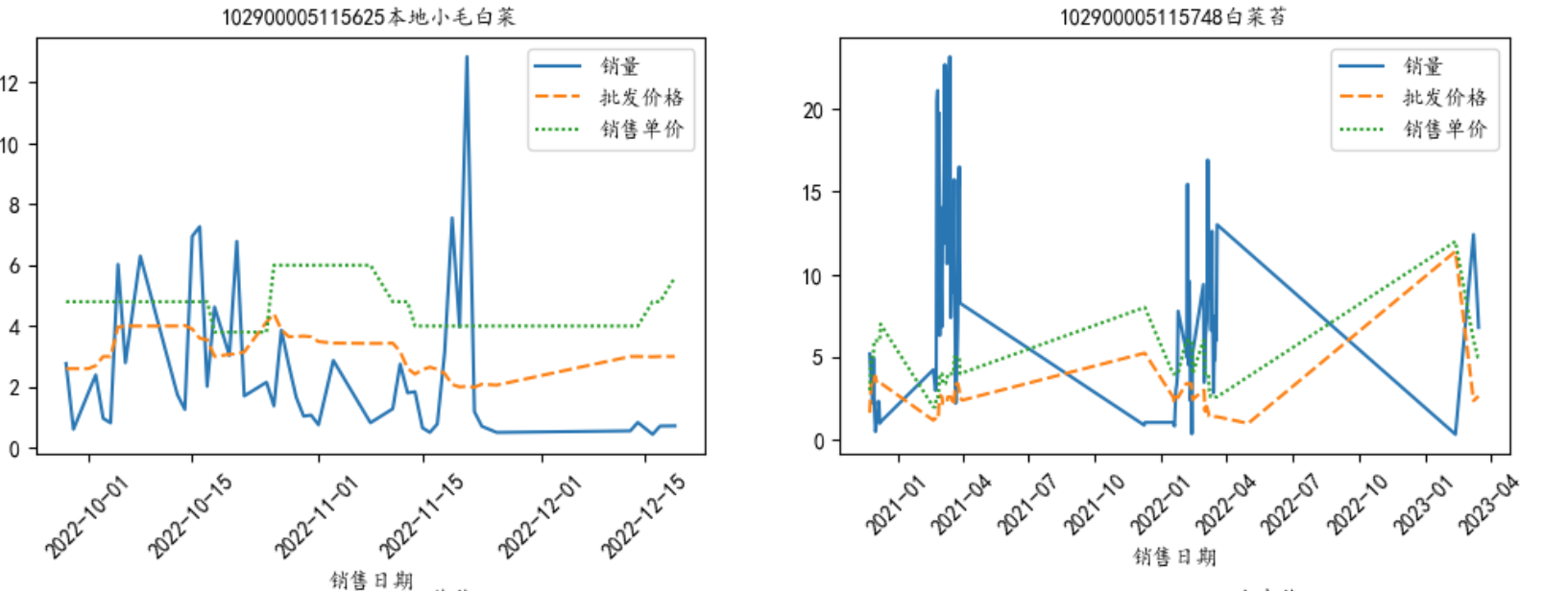
由此可以观察到，大部分品类的销量呈现一定的周期性，说明该部分的品类为季节性商品，而无明显周期性的商品或由于某些其他原因出现峰值。

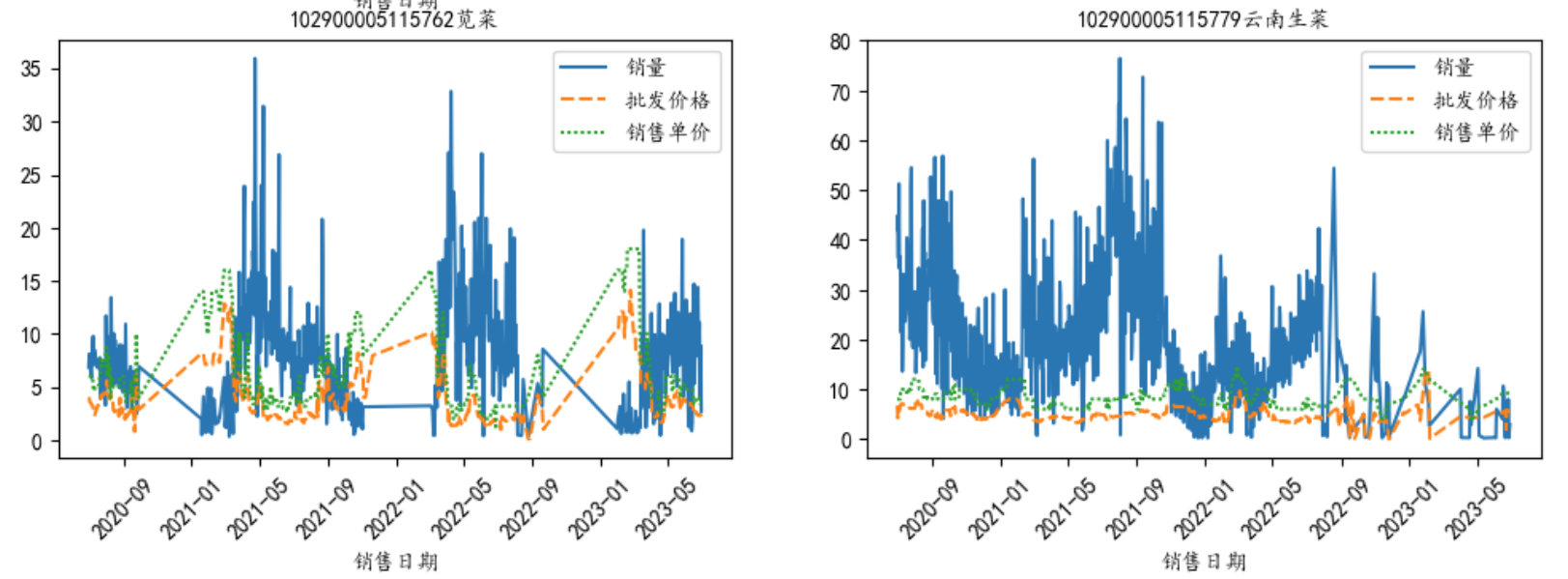
接下来对各商品的销量、销售单价、批发价格进行数据可视化，代码接上一部分如下：

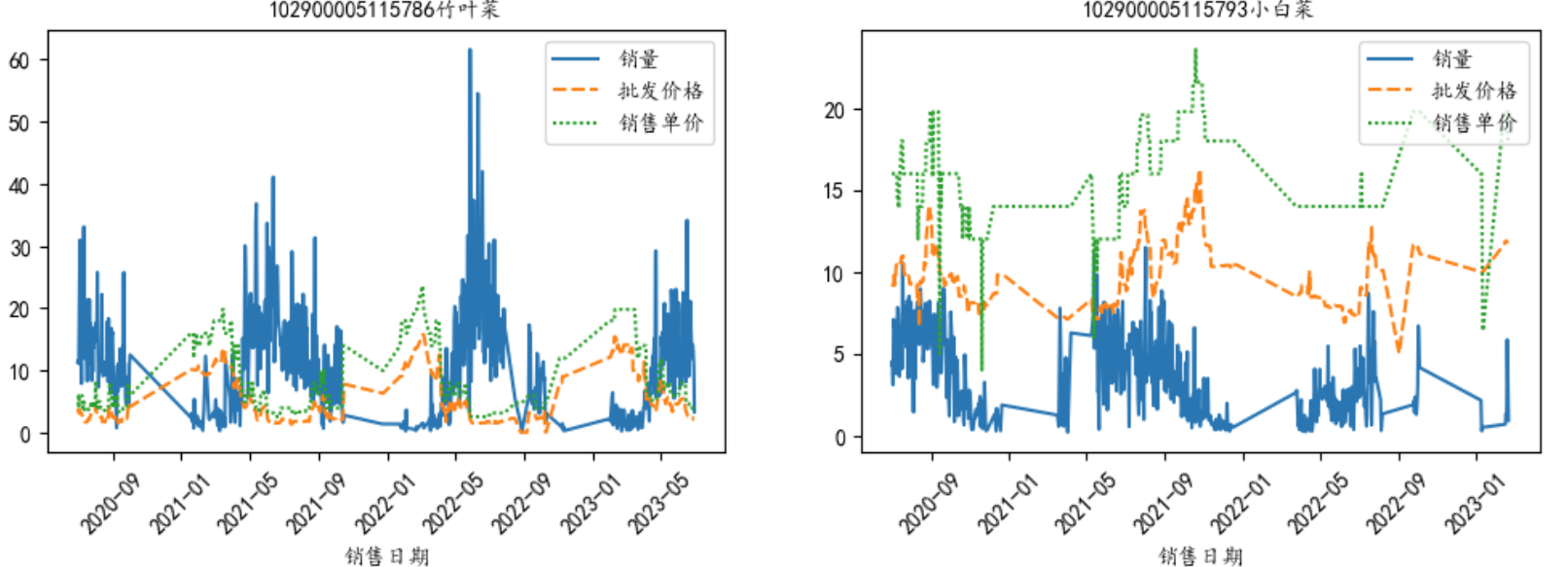


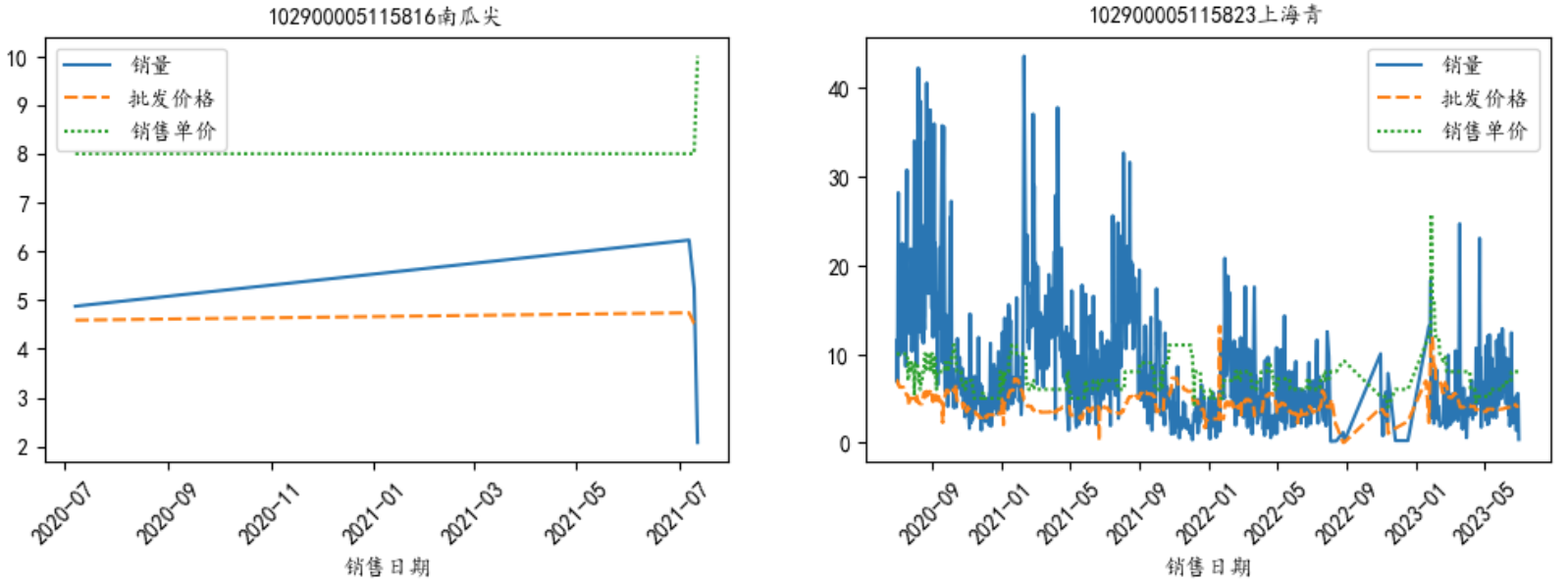
得到的图像如下：

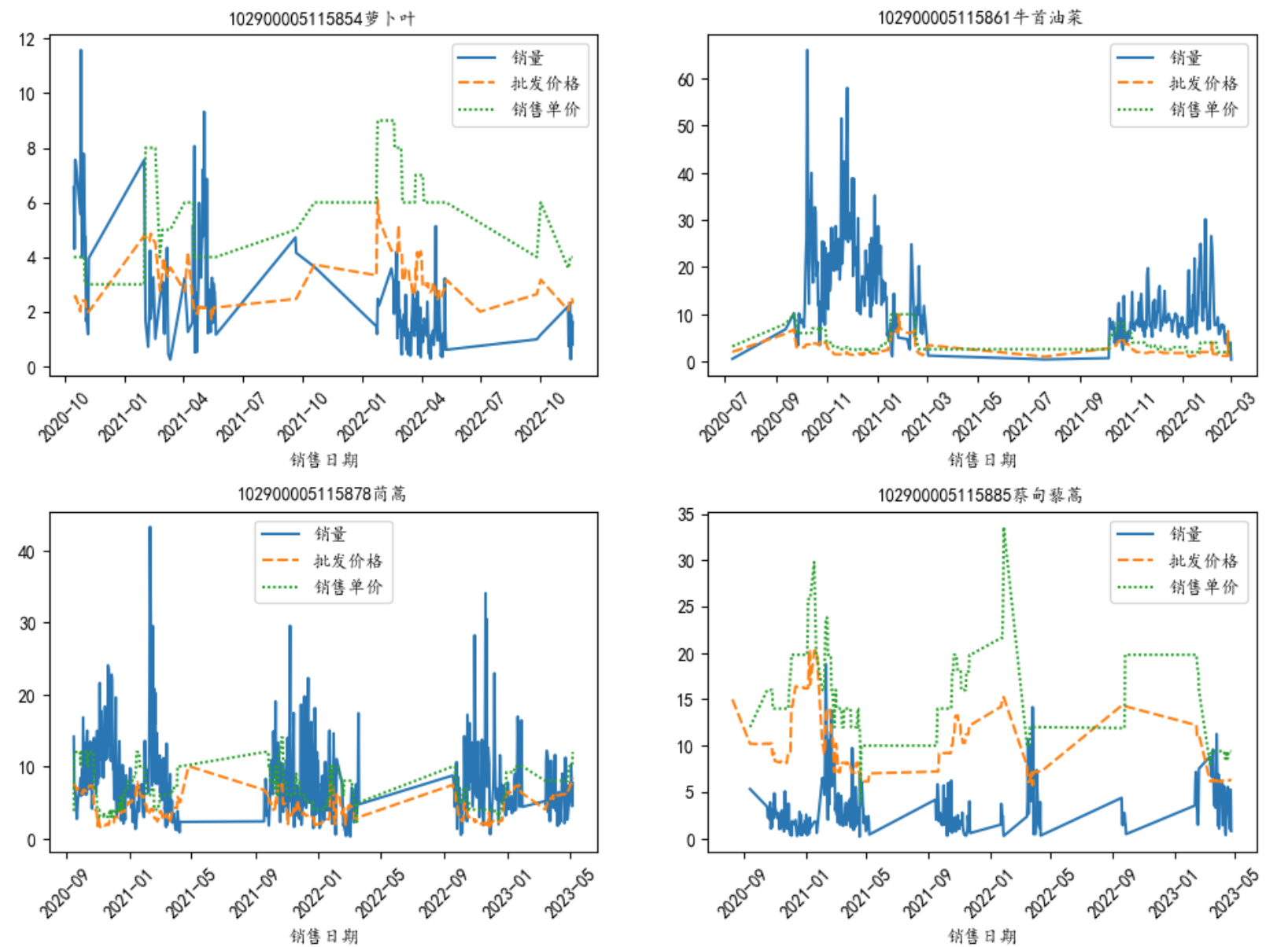












由于图像过多，在此只展示部分图像。

五、数据总结

在利用Pandas + Matplotlib + Seaborn进行数据处理以及可视化后可以看出，销售单价的折线相对批发价格的有一定滞后性，即说明在批发价格上升（降低）后，销售单价也会有一定的提升（降低），具体幅度与商品的种类有关；相似的可以得出，在销售单价提升（降低）后，销量也会呈现一定的降低（上升）。

同时，根据各个商品图像的销量曲线可得到哪些商品为季节性商品。

六、学习总结

通过借助Pandas + Matplotlib + Seaborn这些工具，并以数学建模竞赛中的C题为示例进行学习，我对数据分析有了一定的了解和掌握，相信这项技能会学本人的专业在未来的研究上有一定的帮助。