数据分析报告

侯奕安

目录

1 数据预处理和分析

1.1 数据预处理

1.1.1 异常值检测

①将支付时间转为标准时间的过程中发生错误，经排查错误数据为‘2017/2/29’,后将其修改为‘2017/2/27’。

②经检测发现部分订单应付金额与实付金额都为 0，抹去这部分异常数据。

③在检测过程中发现部分订单中商品金额异常，但由于不确定是否进行了调价或促销，所以并未清除这部分异常订单。

1.1.2 属性修改

①提取售卖机设备 id 后五位，方便后续处理。

②并不清楚后续任务是否需要表中部分属性信息。所以并未删除表中属性。

	订单号	设备ID	应付金额	实际金额	商品	支付时间	地点	状态	提现
0	DD20170816749366361849990978 4	07631	4.5	4.5	68g好丽友巧克力派2枚	2017/1/1 00:53	D	已出货未退款	已提现
1	DD20170816749366355581406116 4	04172	3.0	3.0	40g双汇王米热狗肠	2017/1/1 01:33	A	已出货未退款	已提现
2	DD20170816749357852689093988 6	06874	5.5	5.5	430g泰奇八宝粥	2017/1/1 08:45	E	已出货未退款	已提现
3	DD20170816749368350718661583 7	04228	5.0	5.0	48g好丽友薯原香烤原味	2017/1/1 09:05	C	已出货未退款	已提现
4	DD20170816749375954861825200 6	04134	3.0	3.0	600ml可口可乐	2017/1/1 09:41	B	已出货未退款	已提现
5	DD2017081016294251D0FA5D314F1	04134	4.5	4.5	营养快线	2017/1/1 09:41	B	已出货未退款	已提现
6	DD20170816749366353458905087 1	04228	7.0	7.0	330ml伊利畅意乳酸菌原味	2017/1/1 10:02	C	已出货未退款	已提现

图 1-1-1 异常值监测和属性修改后部分数据

1.1.3 合并信息

以商品名称作为关键词，将附件二中商品信息添加到附件一中。添加后如下图所示。

	订单号	设备ID	应付金额	实际金额	商品	支付时间	地点	状态	提现	大类	二级类
0	DD20170816749366361849909784	07631	4.5	4.5	68g好丽友巧克力派2枚	2017/1/1 00:53	D	已出货未退款	已提现	非饮料	饼干糕点
1	DD20170816749368329675770932	07631	4.5	4.5	68g好丽友巧克力派2枚	2017/1/2 20:58	D	已出货未退款	已提现	非饮料	饼干糕点
2	DD20170816749300229112037656709	06874	4.5	4.5	68g好丽友巧克力派2枚	2017/1/3 01:53	E	已出货未退款	已提现	非饮料	饼干糕点
3	DD20170816749352984906851 4902	04228	4.0	4.0	68g好丽友巧克力派2枚	2017/1/8 20:22	C	已出货未退款	已提现	非饮料	饼干糕点
4	DD20170816749387635309190 9391	04228	14.0	14.0	68g好丽友巧克力派2枚	2017/1/9 21:38	C	已出货未退款	已提现	非饮料	饼干糕点
5	DD20170816749368232379530 9392	07631	4.5	4.5	68g好丽友巧克力派2枚	2017/1/15 22:38	D	已出货未退款	已提现	非饮料	饼干糕点

图 1-1-2 合并附件 1,2 后的数据

1.1.4 按售货机提取数据

因不知道每个地点是否只有一台售货机，所以通过循环获得存有售货机设备 ID 的列表，并根据设备 id 进行分组，将每台售货机的销售数据保存至 csv 文件中，文件名分别为‘task1-1A.csv’‘task1-1B.csv’‘task1-1C.csv’ ‘task1-1D.csv’ ‘task1-1E.csv’

1.2 任务 1.2

提取各售货机五月份销售数据，计算各售货机的订单量和交易额最后汇总获得总订单量和交易额。获得结果如下表。

售货机 / 数据类	1A	1B	1C	1D	1E	ALL
交易额	2392.1	5699	3729	3681	3385.1	18886.2
订单量	553	1287	782	860	750	4232

表 1-2-1 各售货机五月销售情况及汇总

由上表可得，B 售货机销售情况最好，A 售货机销售情况最差，C,D,E 售货机的销售情况相似。

1.3 任务 1.3

任务要求计算每台售货机的每个月的每单平均交易额和每个月的日均交易量。

每个月的每单平均交易额：先通过月份进行分组，对每组内交易额进行加和，最后获取魅族交易单数，相除即可。

每个月的日均交易量：先通过月份进行分组，判定若是 1，3，5，7，8，10，12 则除 31，除二月外其他则除 20,2 月则除 28。即可得到每个月的日均交易量。

1A	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
每单平均交易额	3.74	3.09	4.31	3.82	4.33	4.06	4.26	3.32	3.91	3.9	3.86	3.58
日均订单量	8.26	5.04	6.19	14.19	17.84	33.26	10.16	23.03	31.65	38.1	39.03	53.48

表 1-3-1 A 售货机每月的每单平均交易额与日均订单量

1B	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
每单平均交易额	4.68	3.64	3.59	4.16	4.43	3.84	3.93	3.8	4.14	3.68	4.29	4.17
日均订单量	11.42	9.21	11.29	28.87	41.52	83.16	26.16	57	132.9	89.48	161.65	104.9

表 1-3-2 B 售货机每月的每单平均交易额与日均订单量

1C	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
每单平均交易额	4.36	3.83	3.77	4.42	4.77	4.52	4	3.91	4.44	4.29	4.36	3.95
日均订单量	12.13	7.43	8.48	23.61	25.23	60.48	24.55	40.61	53.97	71.19	62.61	76.55

表 1-3-3 C 售货机每月的每单平均交易额与日均订单量

1D	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
每单平均交易额	3.75	3.26	3.61	4.1	4.28	4.08	4.41	3.58	4.14	4.12	4.28	3.67

日均订单量	11.81	6.61	8.55	19.35	27.74	59.65	11.1	31.65	56.13	65.19	65.39	71.19

表 1-3-4 D 售货机每月的每单平均交易额与日均订单量

1E	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
每单平均交易额	4.52	3.86	3.59	4.06	4.51	4.07	4.11	3.36	4.31	4.03	4.48	3.8
日均订单量	10.77	4.07	8.23	14.32	24.19	53.58	15.32	21.45	33.52	50.35	37.39	64.42

表 1-3-5 E 售货机每月的每单平均交易额与日均订单量

2 数据分析与可视化

2.1 任务 2.1

绘制 2017 年 6 月销量前五的商品销量柱状图:先将时间转换为标准格式，再讲时间列换位到索引上，用户输入要绘制那一月的销量柱状图，通过循环遍历得到次月商品名单，创建等长零列表，两列表压缩成字典，依次更新字典中商品销量，根据销量对字典进行排序，获取用户画前几的柱状图，绘图。

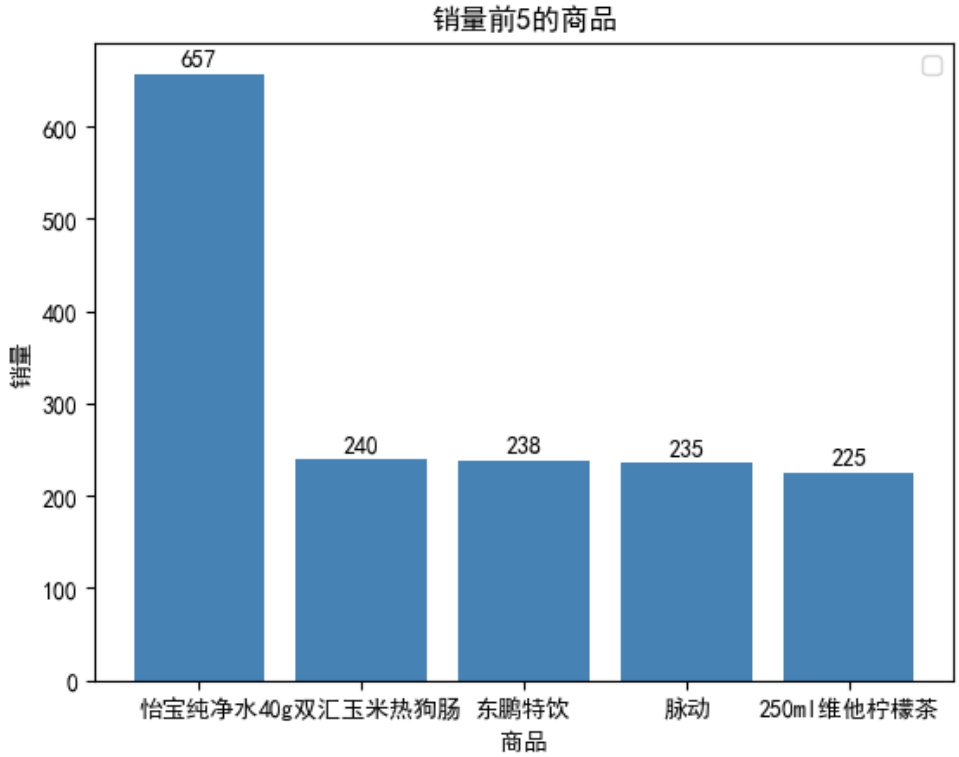


图 2-1-1 六月份销量前五的商品及其销量

2.2 任务 2.2

2.2.1 绘制每台售货机每月总交易额折线图

读取数据后先将支付时间转换为标准时间并换位到索引，新建用于存储总交易额的空列表，通过 resample 和 sum 获得每月的交易额并存除到列表中。新建月份列表用作 x 轴。设定画图参数，画图。

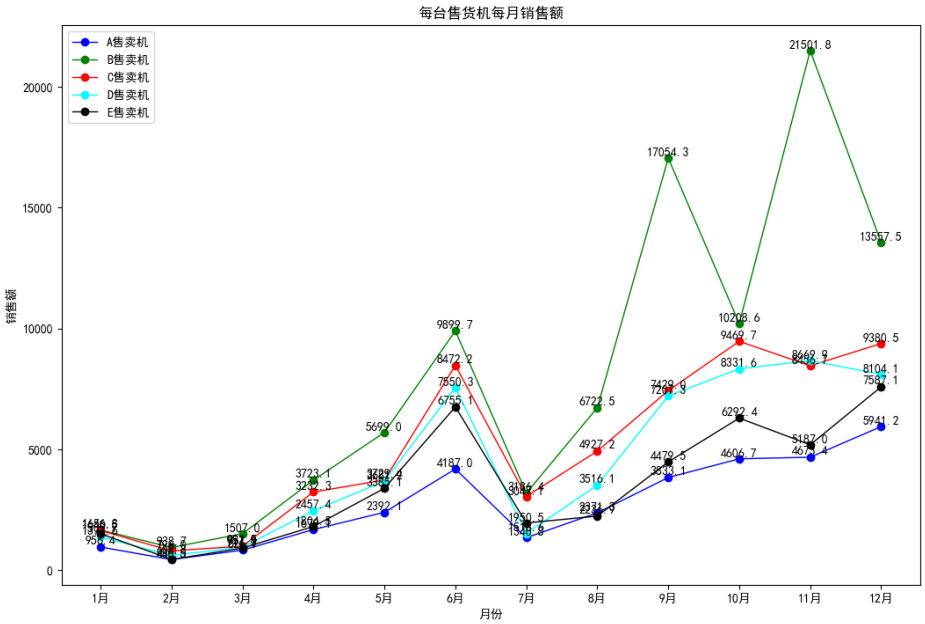


图 2-2-1 每台售货机每月总交易额折线图

由上折线图可得，所有售卖机销售额在整体上都呈上升趋势，且在 6 月出现小高峰，总体上 B 售卖机销售额高于其他售货机。

2.2.2 总交易额月环比增长率

读取数据后先将支付时间转换为标准时间并换位到索引，新建用于存储总交易额的空列表，通过 resample 和 sum 获得每月的交易额并存除到列表中。

通过循环计算所有月环比增长率，并存除到列表中

设定画图参数，画图。

如下列 2-2-2.1,2-2-2.2,2-2-2.3,2-2-2.4,2-2-2.5 五张图所示：

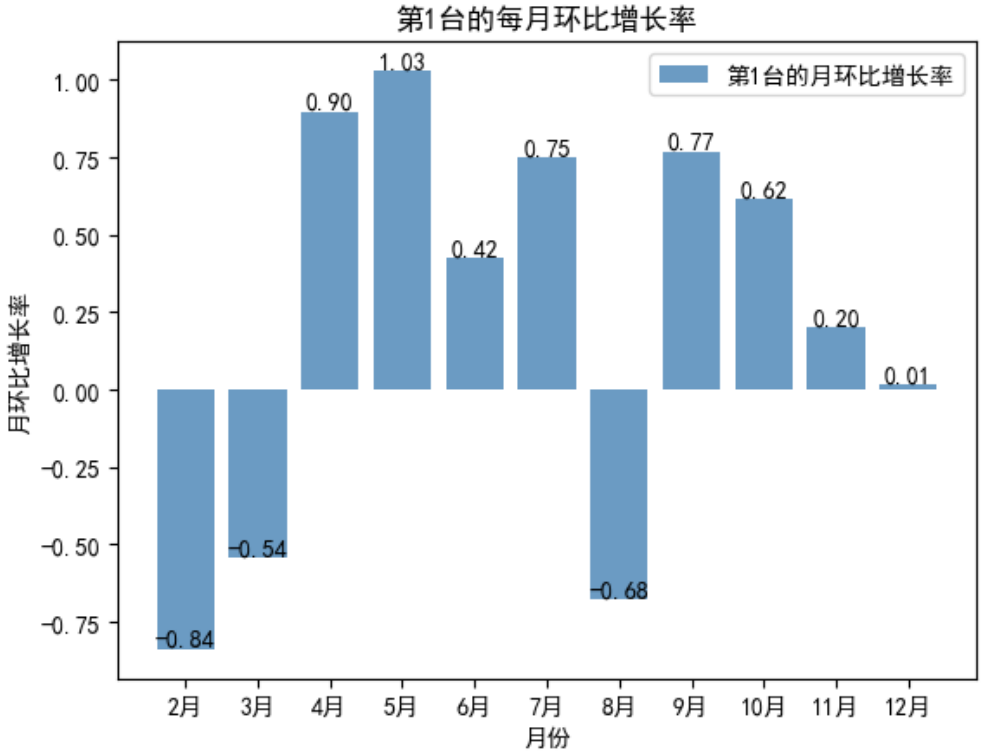


图 2-2-2.1 第一台售货机每月环比增长图

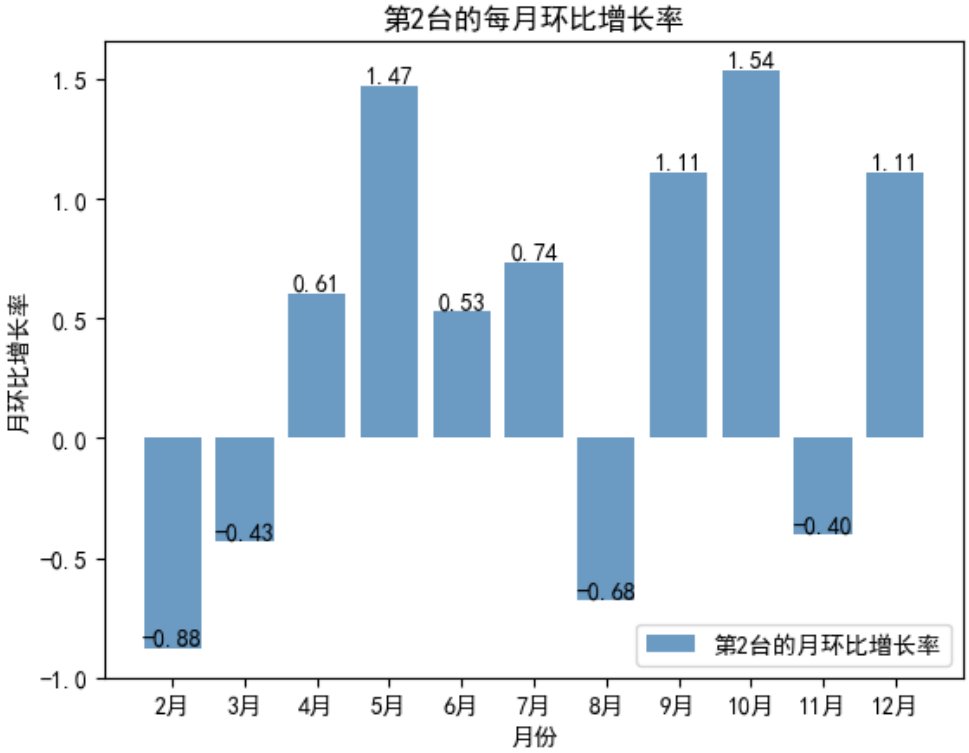


图 2-2-2.2 第二台售货机每月环比增长图

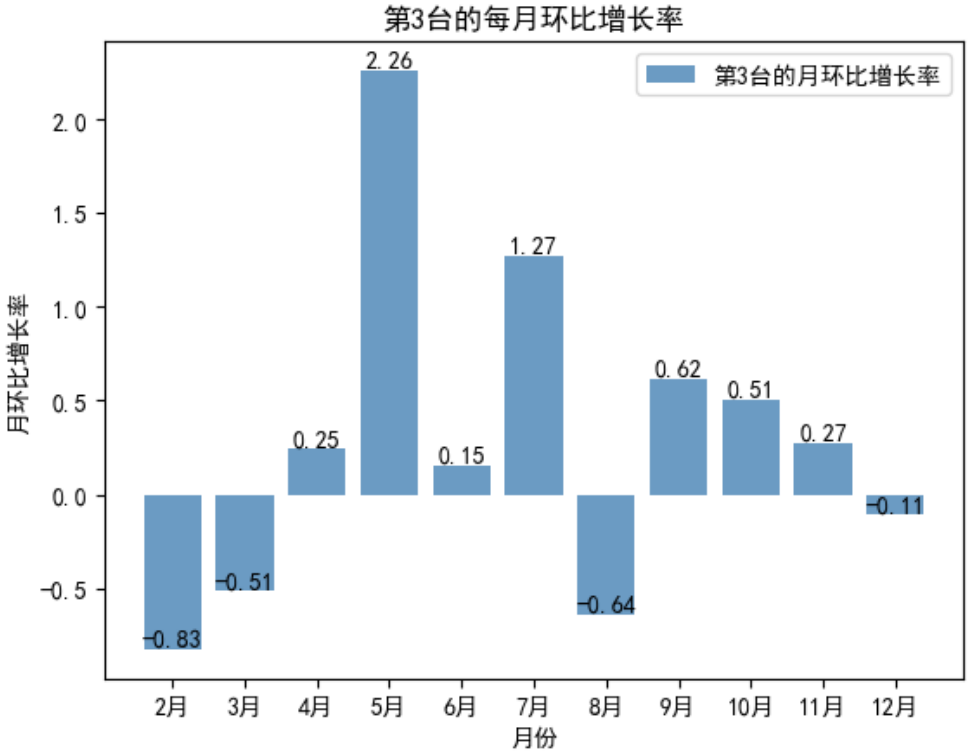


图 2-2-2.3 第三台售货机每月环比增长图

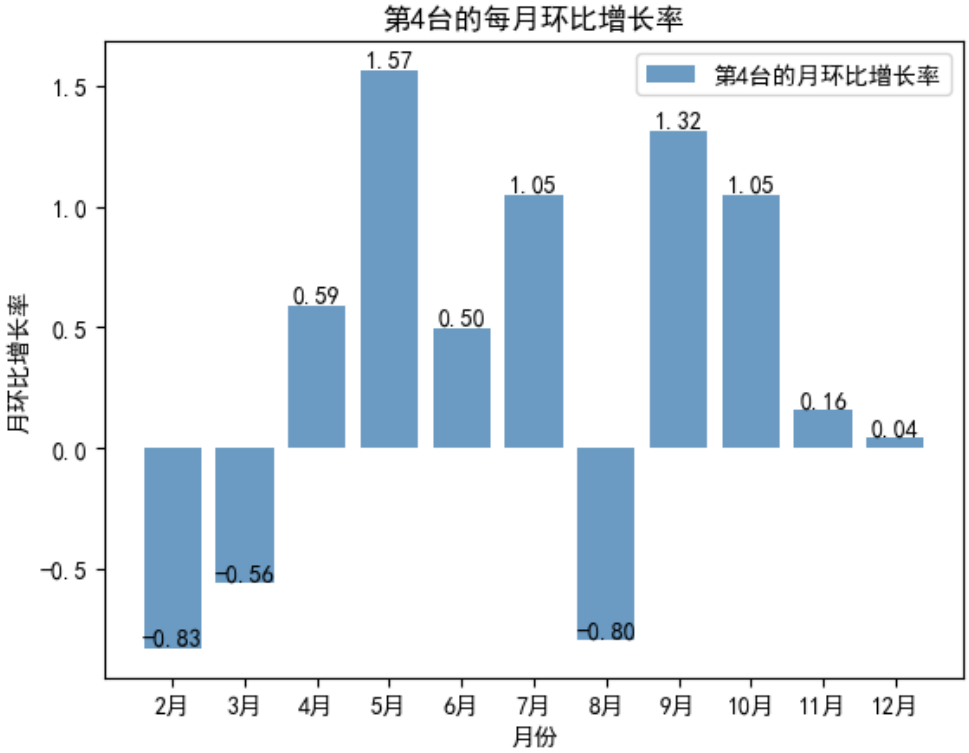


图 2-2-2.4 第四台售货机每月环比增长图

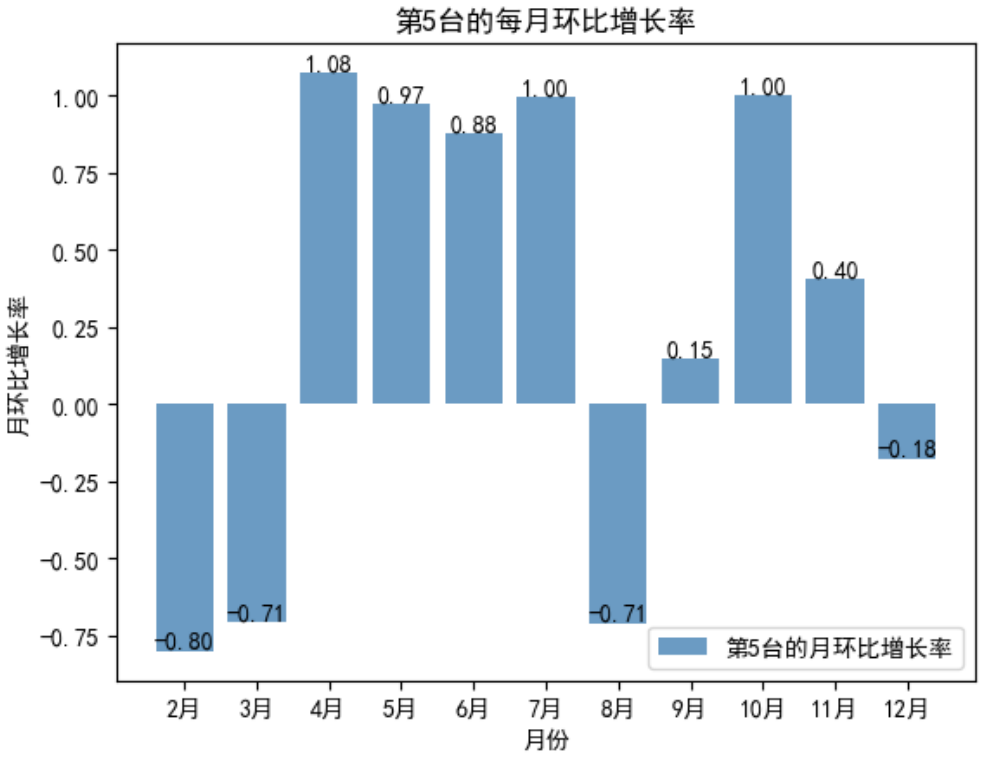


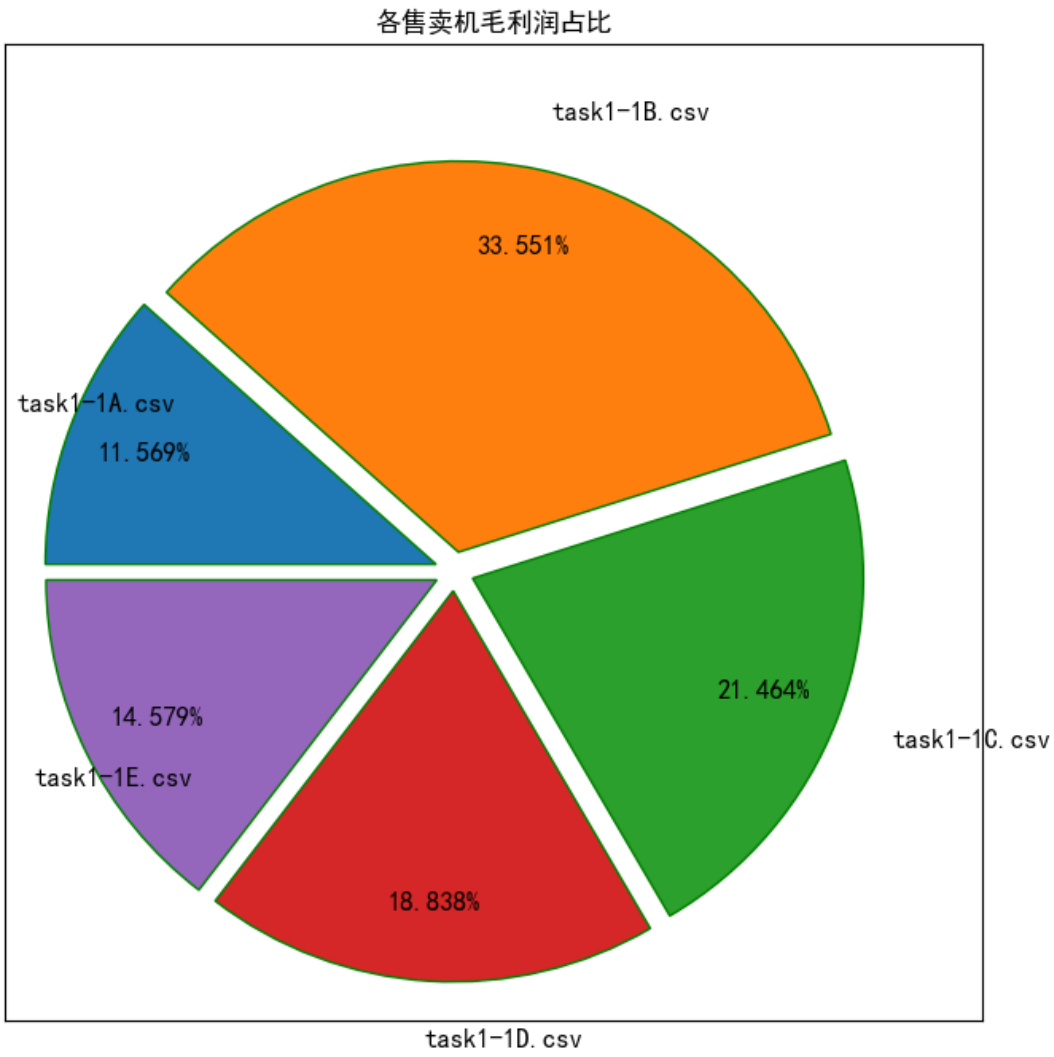
图 2-2-2.5 第五台售货机每月环比增长图

从上方五张月环比增长图来看，每个售货机在 2,3,8 月都出现了负增长的情况，在第二台售货机的 11 月也出现了负增长的情况。在 A 售货机中，增长的月环比整体呈一个下降趋势，在 B 售货机中，增长的月环比整体呈先升后降趋势，在 C 售货机中，增长的月环比整体呈一个下降趋势，在 D 售货机中，增长的月环比整体呈起伏趋势，在 E 售货机中，增长的月环比整体呈一个平稳趋势。

2.3 任务 2.3

各售货机毛利润站总毛利润比例饼图：现在附加二中读取分类标准，将饮料类存放在饮料类列表中，非饮料类存放在非饮料类列表中，通过循环和判断的结合，如果商品在饮料类中则应付金额*0.25，如果商品在非饮料类中则应付金额*0.2。将五个售货机的毛利润放在列表中设置参数进行画图。

如下图所示：

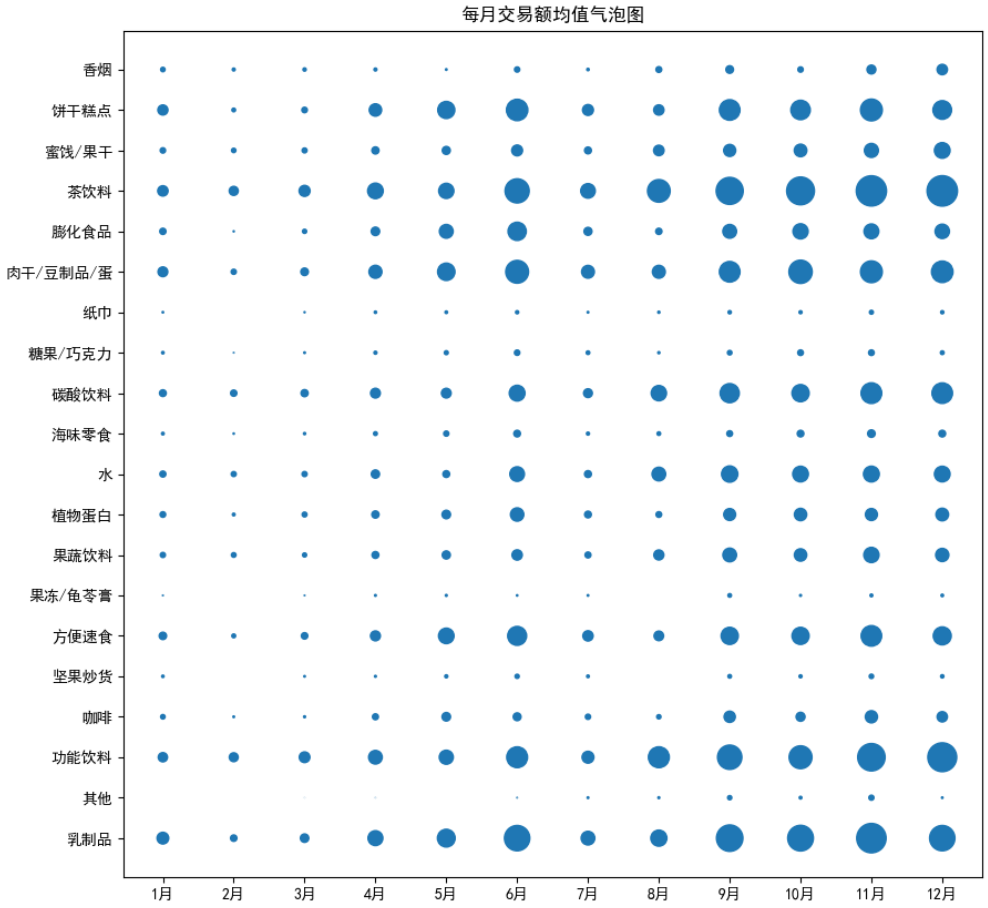


如图所示售卖机 B 毛利润所占比例最大，占比为 33.55%，A，E 售卖机毛利润所占比例最少，A 为最低只占 11.57%，整体呈 BCDEA 依次下降趋势。

2.4 任务 2.4

每月交易额均值气泡图：读取数据，将时间调整至标准格式，换位给索引，通过 groupby 和 sum 获得每个月二级类销量，十二个月合并到一起，对空缺值进行处理(设置为 0)，处理列名，组成索引为商品名称，属性为 12 个月份，值为销售额的表。

设置参数，绘图：



由气泡图可得，茶饮料，功能饮料，乳制品从高到低依次占据了交易额均值前三的位置。同时也发现所有商品以半年为一个周期，交易额均值每个月依次增加。

2.5 任务 2.5

绘制售货机 C6,7,8 三个月订单的热力图：读取数据，支付时间调整至标准格式并设为索引

通过分组获取确定日期销量，并放入矩阵对应位置中，将矩阵转换为 dataframe 格式，更新索引和属性。设置画图参数并画图。

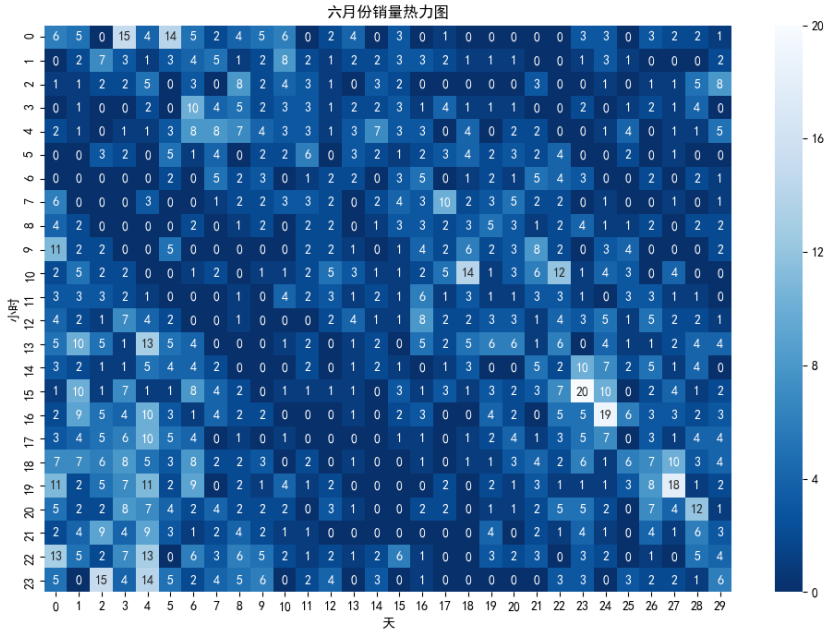


图 2-5-1 C 售货机 6 月份订单量热力图

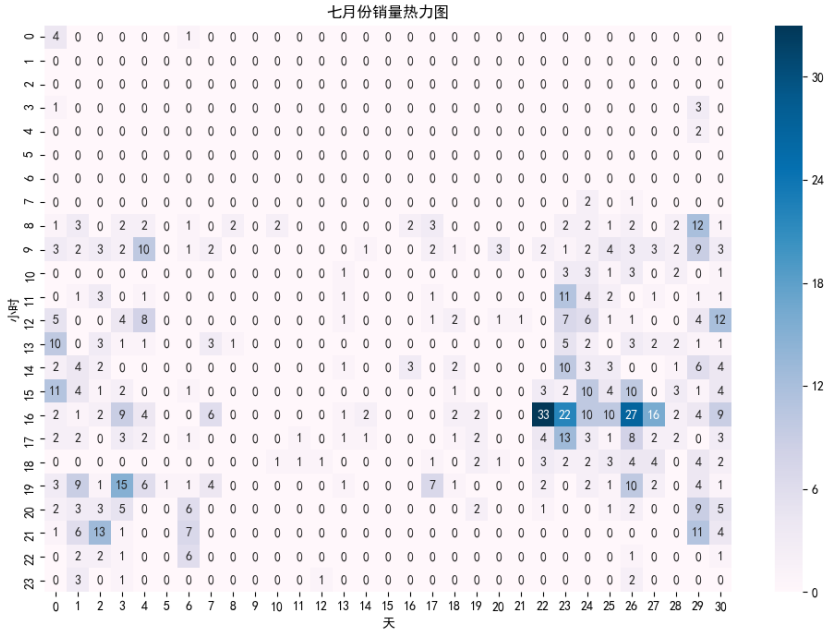


图 2-5-2 C 售货机 7 月份订单量热力图

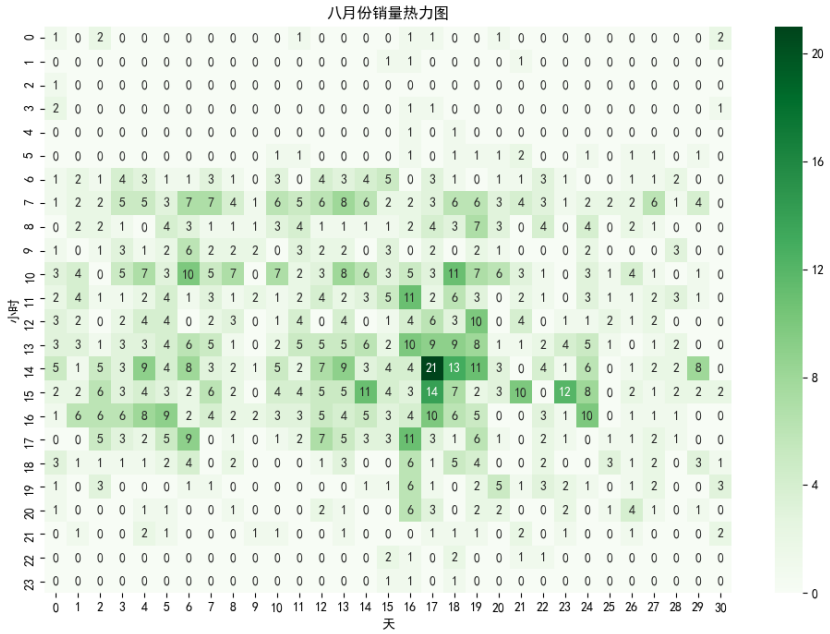


图 2-5-3 C 售货机 8 月份订单量热力图

从上图可得，在六月份时销售基本集中在上旬和下旬，中旬占小部分，交易时间集中于下午，16 点左右。

在七月份时销售基本集中在下旬和上旬，中旬占小部分，交易时间集中于下午，16 点左右。

在八月份时销售基本集中在中旬，中旬占小部分，交易时间集中于下午，16 点左右。

由此可得，在六七八月时，人们通常在八点以后才进行购物活动，可以赶在八点之前进行补货，保证销售供应。同时在下午 16 点左右会迎来销售高峰，所以赶在 16 点之前进行检查，对缺货商品进行补货。

3 自动售货机画像

3.1 贴标签

设定评价指标=销量*0.6+毛利润*0.4

给所有商品贴标签：读取数据，通过商品名进行分类，并得出统计数量作为销量，得到销量。听过饮料与非饮料类，通过商品计算获得毛利润，将销量和毛利润通过商品合并入总表中。计算获得评价指标。将评价指标由高到低降序排列，取排名前 5%作为热销类商品，5%-70%作为正常销售类商品，剩余为滞销类商品，

将类别存入销售情况列中。

3.2 画像

通过扩展后的标签生成画像。取评价指标前 50 个进行画像

3-3-1 A 售货机画像

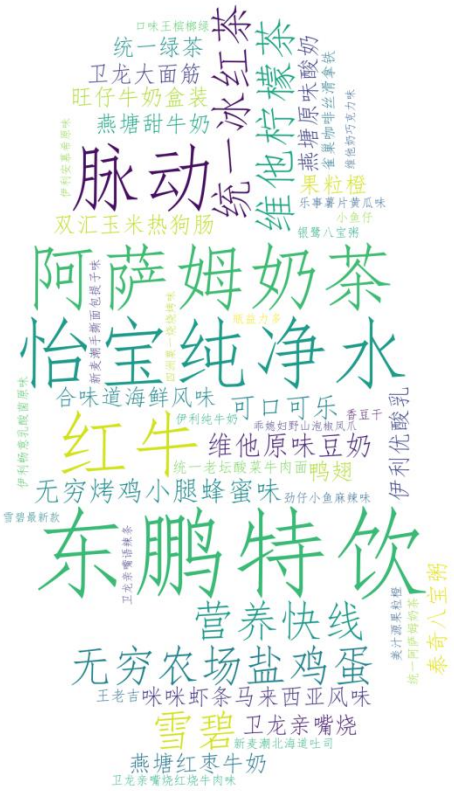

脉动
营养快线
怡宝纯净水
东鹏特饮
阿萨姆奶茶
红牛
统一冰红茶
维他原味豆奶
维他柠檬茶
统一绿茶
果粒橙
双汇玉米热狗肠
无穷农场盐鸡蛋
燕塘甜牛奶
燕塘原味酸奶
银鹭八宝粥
可口可乐
雪碧
王老吉罐
伊利优酸乳
伊利纯牛奶
咪咪虾条马来西亚风味
无穷烤鸡小腿蜂蜜味
香豆干
鸭翅
芬达
合味道海鲜风味
泰奇八宝粥
燕塘红枣牛奶
雪碧最新款
可口可乐最新款
劲仔小鱼麻辣味
维他奶巧克力味
统一老坛酸菜牛肉面

3-3-2 B 售货机画像


伊利优酸乳
营养快线
脉动
咪咪虾条马来西亚风味
合味道海鲜风味
统一绿茶
泰奇八宝粥
红牛
怡宝
王老吉
旺仔牛奶盒装
果粒橙
无穷烤鸡小腿蜂蜜味
统一冰红茶
东鹏特饮
无穷农场盐鸡蛋
芬达
鸭翅
雪碧最新款
燕塘红枣牛奶
伊利纯牛奶
王老吉罐
怡宝纯净水
维他原味豆奶
燕塘甜牛奶
燕塘原味酸奶
维他柠檬茶
双汇玉米热狗肠
阿萨姆奶茶
可口可乐
银鹭八宝粥

3-3-3 C 售货机画像

3-3-4 D 售货机画像



3-3-5 E 售货机画像

由上图可见，“东鹏特饮”，“怡宝纯净水”，“营养快线”“阿萨姆奶茶”等销量在 ABCDE 售货机上得评价都较为突出，应加大这部分的商品的供应量，以保证不缺货。

4 业务预测

4.1 预测原理与能否通过已有数据进行预测的原因

由于已有数据实在是太少了，我不认为可以在此基础上可以得到较为良好的预测模型。

即使我认为不能得到良好的模型，但还是要预测试验一下，考虑到数据量极其少，所以我先用了对小样本较好的 SVM 进行回归并预测。

预测原理：

支持向量机目标：所有误分类的点(Mi)到超平面的距离和最小，即下式最小

$$\sum_{x\in Mi}\frac{-y^{(i)}(w^Tx^{(i)}+b)}{||w||_2}$$

当 w 和 b 成比例增加时，分母的 L2 番薯也会成比例增加，由此分子和分母有固定的倍数关系，所以我们固定分子或分母为 1，然后求另一个即分子自己或分母的倒数的最小化作为损失函数来简化损失函数，我才用保留分子，固定分母 $||w||_2=1$，此时简化后最终的感知机模型的损失函数：

$$\sum_{x\in Mi}\frac{-y^{(i)}(w^Tx^{(i)}+b)}{1}$$

SVM 的模型是让所有点到超平面的距离大于一定距离，也就是所有的分类点要在各自类别的支持向量两边。即：

$$\max\gamma=\sum_{x\in Mi}\frac{y^{(i)}(w^Tx^{(i)}+b)}{||w||_2}s.t\ y_i((w^Tx_i+b))=\gamma`^{(i)}\geq\gamma`(i=1,2\dots m)$$

一般取函数间隔$\gamma`=1$,所以优化函数为：

$$\max\frac{1}{||w||_2}s.t\ y_i((w^Tx_i+b))\geq 1(i=1,2\dots m)$$

也就是说要最大化$\frac{1}{||w||_2}$，又因为$\max\frac{1}{||w||_2}=min||w||_2{}^2$，所以：

$$min\frac{1}{2}||w||_2{}^2 s.t\ y_i((w^Tx_i+b))\geq 1(i=1,2\dots m)$$

所以至此，我们的目标函数是让权值的二范数最小，同时让各个训练数据尽量远离自己类别已变得支持向量即：

$$y_i(w*\varphi(x_i)+b)\geq 1$$

加入松弛变量$\xi_i \geq 0$，则目标函数为：

$$\frac{1}{2}||w||_2{}^2+C\sum_{i=1}^{m}\xi_i \quad ,y_i(w*\varphi(x_i)+b)\geq 1-\xi_i$$

我们定义 ϵ>0,对于点(x_i,y_1)有：

$$\begin{cases} \text{损失为}0, & |y_i(w*\varphi(x_i)+b)|\leq\epsilon \\ \text{损失为}|y_i(w*\varphi(x_i)+b)|-\epsilon, & |y_i(w*\varphi(x_i)+b)|>\epsilon \end{cases}$$

所以 SVR 目标函数的形式为

$$min\frac{1}{2}||w||_2{}^2+C\sum_{i=1}^{m}\xi_i^{\wedge}+\xi_i^{\vee}$$

$$s.t.-\epsilon-\xi_i^{\vee}\leq y_i-w*\varphi(x_i)-b\leq\epsilon+\xi_i^{\wedge}$$

$$\xi_i^{\wedge}\geq 0,\xi_i^{\vee}\geq 0(i=1,2,\ldots.,m)$$

4.2 预测结果

对数据的要求：最好多给几年的数据，只有一年的数据误差过于巨大。

预测结果：

A	预测销售额	均方误差	解释方差	可决系数
饮料	911.7	194717	0.8616	0.8609
非饮料	476.7	281774	0.37	0.3314
B	预测销售额	均方误差	解释方差	可决系数
饮料	2182	8250885	0.6096	0.5642
非饮料	455.8	2369291	0.5355	0.4398
C	预测销售额	均方误差	解释方差	可决系数
饮料	1452	1297623	0.7408	0.707
非饮料	721	1010230	0.3505	0.2915
D	预测销售额	均方误差	解释方差	可决系数
饮料	1329.8	1174716	0.7138	0.7103
非饮料	503	946773	0.2586	0.2554

E	预测销售额	均方误差	解释方差	可决系数
饮料	717.4	717113	0.643	0.6027
非饮料	780	630190	0.4312	0.4187

下列图为对比图(可若看不清可将图拖大)：

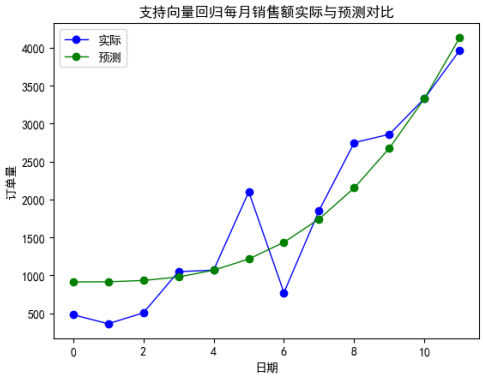


图 4-2-1 A 售货机饮料类预测

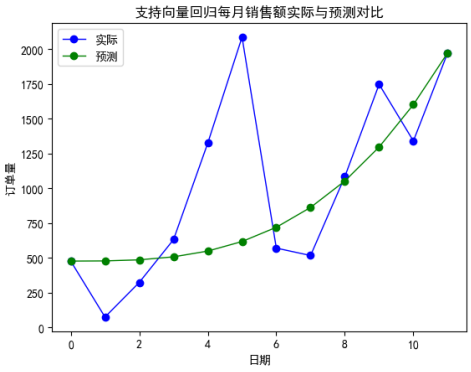


图 4-2-2 A 售货机非饮料类预测

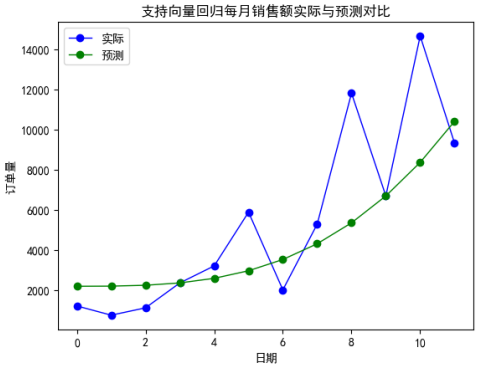


图 4-2-3 B 售货机饮料类预测

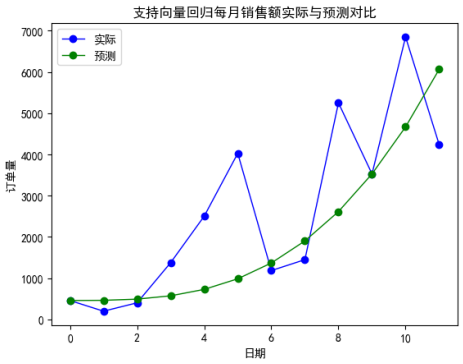


图 4-2-4 B 售货机非饮料类预测

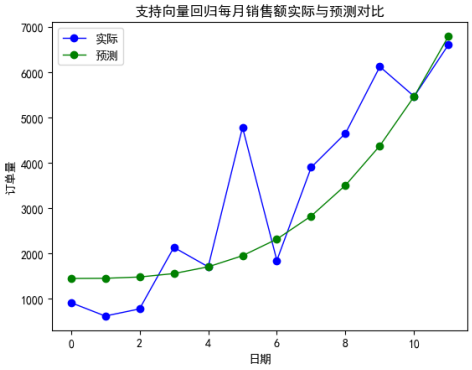


图 4-2-5 C 售货机饮料类预测

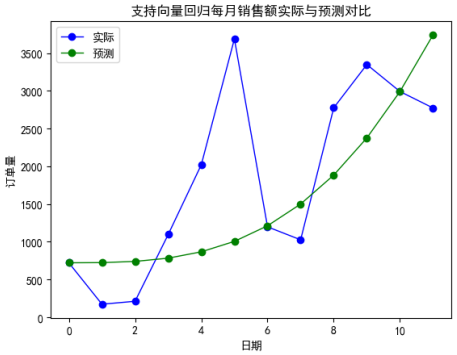


图 4-2-6 C 售货机非饮料类预测

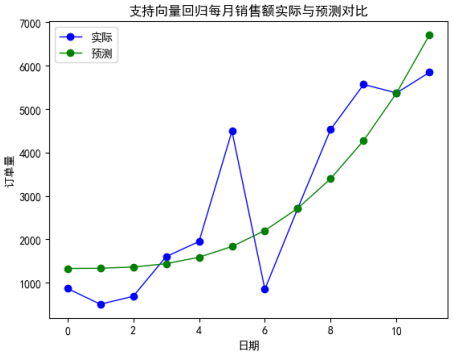


图 4-2-7 D 售货机饮料类预测

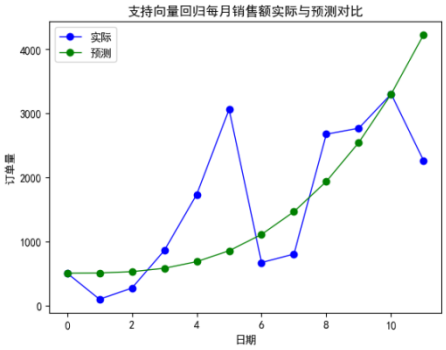


图 4-2-8 D 售货机非饮料类预测

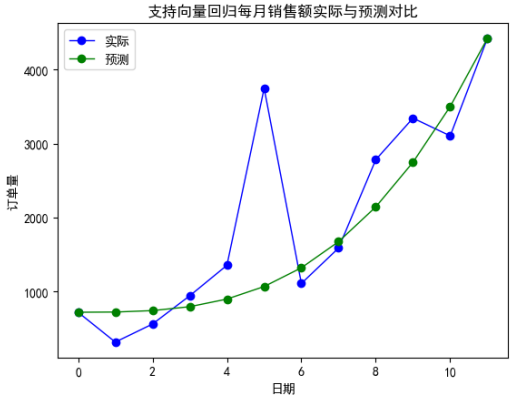


图 4-2-9 E 售货机饮料类预测

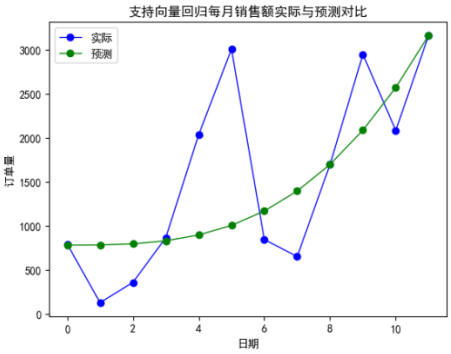


图 4-2-10 E 售货机非饮料类预测