新零售——无人智能售货机

商务数据分析报告

姓名：周诗玮

联系方式：13679592711

邮箱：372040071@qq. com

目录

一、问题重述

1.1 背景

自动售货机以线上经营的理念，提供线下的便利服务，以小巧、自助的经营模式节省人工成本，让实惠、高品质的商品触手可及，成为当下零售经营的又一主流模式。自动售货机内商品的供给频率、种类选择、供给量、站点选择等是自动售货机运营者需要重点关注的问题。因此，科学的商业数据分析能够帮助经营者了解用户需求，掌握商品需求量，为用户提供精准贴心的服务，是掌握经营方向的重要手段，对自动售货机这一营销模式的发展有着非常重要的意义。

某商场在不同地点安放了 5 台自动售货机，编号分别为 A、B、C、D、E。附件 1 提供了从 2017 年 1 月 1 日至 2017 年 12 月 31 日每台自动售货机的商品销售数据，附件 2 提供了商品的分类。

1.2 目标

根据自动售货机的经营特点，对经营指标数据、商品营销数据及市场需求进行分析，完成对销量、库存、盈利三个方面各项指标的计算，按要求绘制对应图表，并预测每台售货机的销售额。

为每台售货机所销售的商品贴上标签，使其能够很好地展现销售商品的特征。

二、数据预处理与分析

导入附件一的数据并进行初步的处理审查。首先检查是否存在缺失值和订单实际金额与应付金额有差异的情况，由图 1 可见，原数据集中不存在缺失值，通过循环遍历比对，亦可知无实际金额与应付金额不一致的情况。

```
In [2]: data.isnull().any()

Out[2]: 订单号      False
        设备ID     False
        应付金额     False
        实际金额     False
        商品       False
        支付时间     False
        地点       False
        状态       False
        提现       False
        dtype: bool
```

图 1 检查缺失值

其次，对支付时间的数据类型进行定义，即将其转化成时间类型数据。该过程出现错误如图 2，由此可推测存在时间异常的数据。这种情况下，将支付时间设置为索引，把数据按顺序排列并检查两端的数据，运行结果如图 3。可见最后一行的支付时间异常，不存在 2017 年 2 月 29 号。遂删除该行数据后，成功完成支付日期数据类型转换。

```
ValueError: day is out of range for month
```

图 2 时间类型数据转换错误

2017/12/31 22:39	DD20170613020607768E3940FA188
2017/12/31 23:10	DD2017060217303716A53CCD6B185
2017/2/29 3:44:00 PM	DD20170816749324155469202675 2

70680 rows × 8 columns

图 3 支付日期异常值

数据的预处理初步完成后，把附件一的各项数据按地区分类，提取每台售货机对应的销售数据保存至”task1-1A.csv”、”task1-1B.csv”、…、”task1-1E.csv”。

接下来，计算每台售货机 2017 年 5 月得到交易额、订单量及所有售货机交易总额和订单总量，结果详见表 1。

表 1 2017 年 5 月交易额及订单量

地区	交易额	订单量
A	3385.1	756
B	3681.2	869
C	3729.4	789
D	2392.1	564
E	5699.0	1292
合计	18886.8	4270

计算每台售货机日均订单量及每月每单平均交易额，结果详见表 2。

表 2 日均订单量与每月每单平均交易额

	A		B		C		D		E	
	订单量	交易额	订单量	交易额	订单量	交易额	订单量	交易额	订单量	交易额
1 月	10	4.51	11	3.75	12	4.33	8	3.69	11	4.68
2 月	4	3.86	6	3.26	7	3.83	5	3.09	9	3.64
3 月	8	3.59	8	3.61	8	3.77	6	4.31	11	4.31
4 月	14	4.04	20	4.08	24	4.40	14	3.79	29	4.16
5 月	24	4.48	28	4.24	25	4.73	18	4.24	41	4.41
6 月	55	4.05	61	4.07	62	4.50	34	4.03	86	3.82
7 月	15	4.10	11	4.40	24	3.99	10	4.23	26	3.92
8 月	21	3.36	31	3.58	40	3.91	23	3.32	57	3.80
9 月	34	4.31	58	4.13	55	4.43	32	3.90	137	4.13
10 月	50	4.02	65	4.11	71	4.27	38	3.90	89	3.68
11 月	38	4.47	67	4.26	64	4.35	40	3.86	167	4.28
12 月	64	3.79	71	3.67	76	3.94	53	3.57	104	4.17

由表 2 可见，各地区每月每单平均交易额相差不大，均分布在 3-5 元之间，而日均订单量则有较大差异，E 地区 11 月份日均订单量可达最大值 167 个/天，而 A 地区 2 月份日均订单量仅为 4 个/天，为最小值。故可推测，不同地区的人群对自动售货机的偏好有显著差异，E 地区的人群对自动售货机的商品需求量最大。由于每单交易额没有显著性差异，因此各地区售货机的总交易额的差异主要是由订单量差异导致，我们可以针对性地对需求量不同的地区采用不同的进货、存货策略。

三、数据分析与可视化

3.1 2017 年 6 月销量前五商品柱状图

读取预处理后的数据，使用聚合函数获取订单量前五的商品名称及销量，获取结果如图 3。

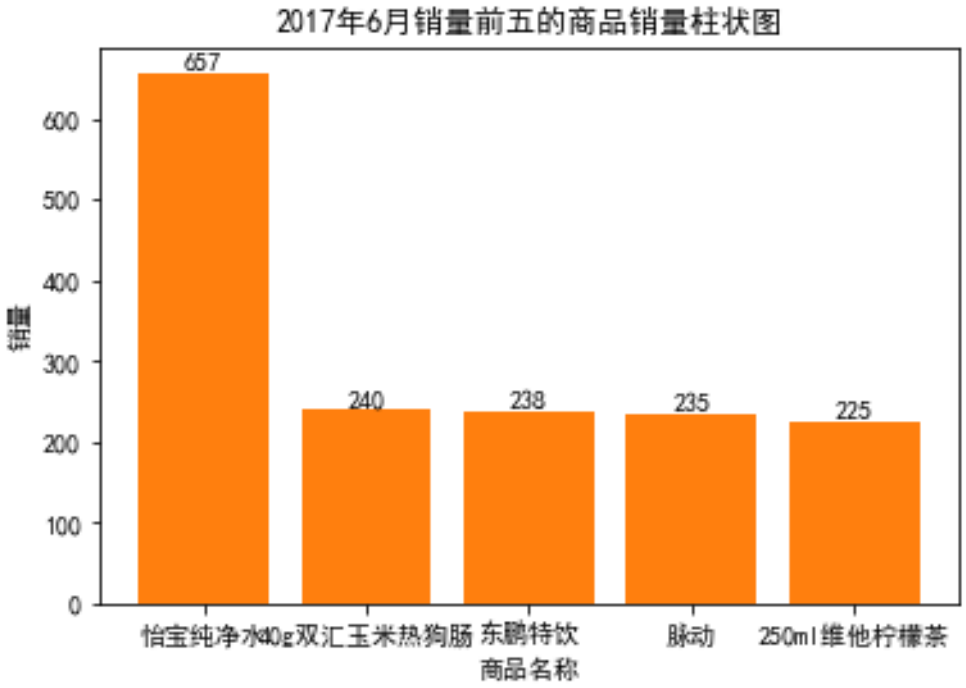

图 4 2017 年 6 月销量前五商品

分析可得，2017 年 6 月销量前五的商品均为饮料类商品，其中怡宝纯净水销量最高，远超其余四种饮料，其余四种饮料的销量差异相对较小。因此可推测，人们对自动售货机的饮料类商品的需求比非饮料类商品高，且在饮料类商品中，纯净水需求量最大，可适当提高各地区纯净水的进货量。

3.2 各地区每月总交易额折线图及月环比增长率柱状图

为了获取月交易总额数据，设置索引为支付时间，按月计算交易总额，并计算交易额月环比增长率

$$\frac{M(i)-M(i-1)}{M(i-1)}$$ ，

其中 $M(i)$为第 i 期交易额。获取数据后绘制每个地区每月销售总交易额折线图及环比增长率柱状图如图 5-9。

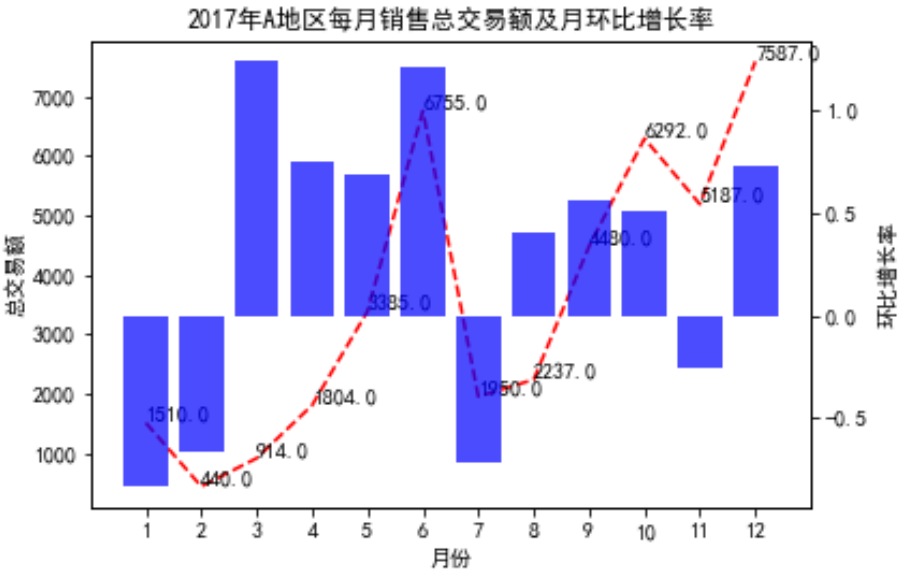

图 5 A 地区每月总交易额折线图及月环比增长率

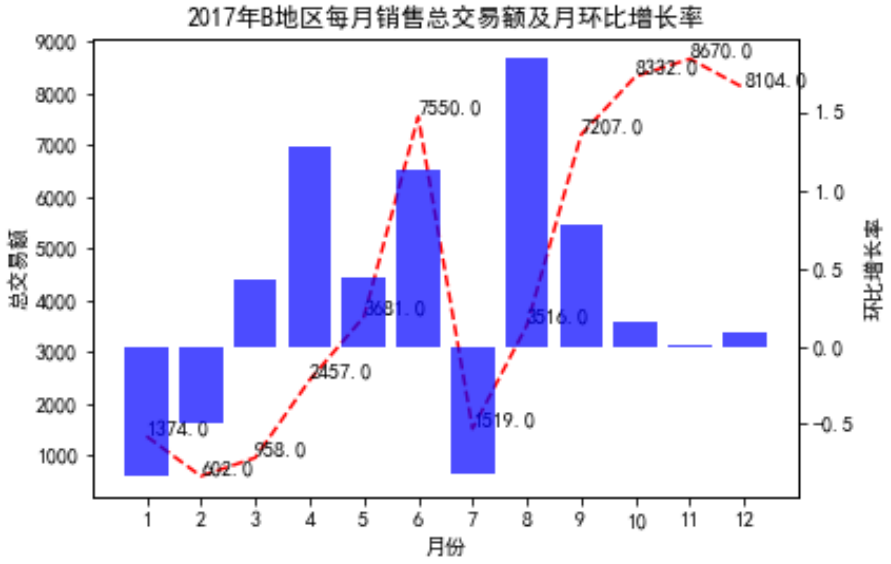

图 6 B 地区每月总交易额折线图及月环比增长率

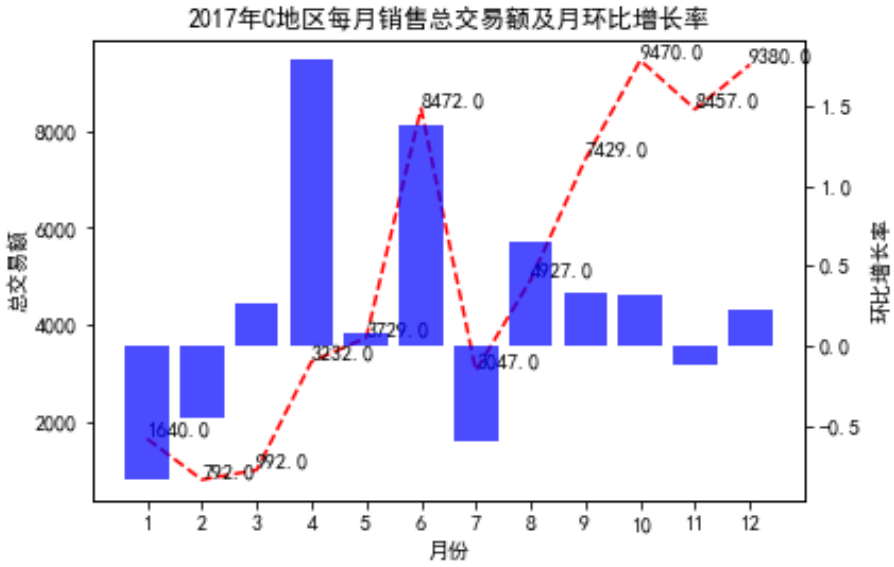

图 7 C 地区每月总交易额折线图及月环比增长率

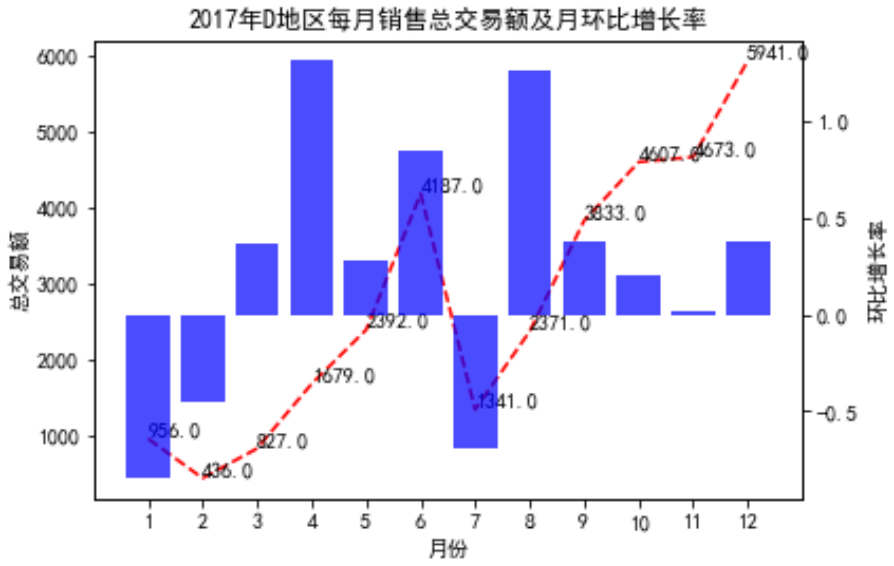

图 8 D 地区每月总交易额折线图及月环比增长率

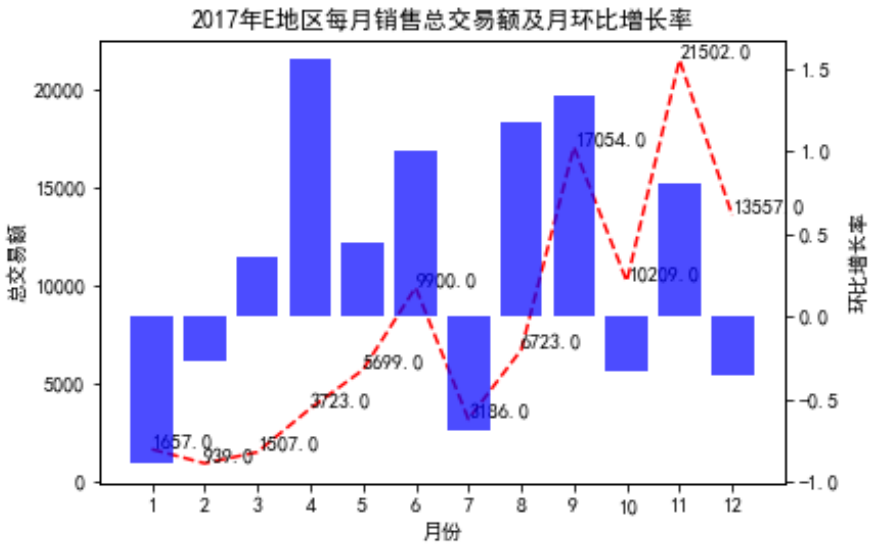

图 9 E 地区每月总交易额折线图及月环比增长率

分析可得，各地区的交易总额变化趋势大致相同，1 月开始交易额逐步上升，在 7 月骤降，然后 7-12 月交易额逐步回升。总体而言，2017 年上半年人们对自动售货机商品需求量较小，年末各地区人群的需求量均达到较高水平。排除季节因素影响，导致人们需求量变化的原因可能是人们对自动售货机的了解程度及使用习惯发生了改变。2017 年初，各地区的自动售货机刚刚部署，人们对此知之甚少，因此售货机销量额普遍不高；随着时间推移，越来越多的人开始使用自动售货机，售货机凭借其便利性吸引了越来越多的顾客。

3.3 各地区毛利润占比饼图

整合附件一和附件二至”3.csv”，根据商品大类数据对商品进行饮料类和非饮料类分类，计算分类后的交易总额。假定饮料类商品的毛利率为 25%，非饮料类商品的毛利率为 20%，计算各地区毛利润占总毛利润的比例，绘制饼图如图 10。

据图可见，E 地区毛利润占比最大，其次是 C 地区、B 地区和 A 地区，D 地区所获毛利润占比最小。针对各地区毛利润的差异，可采取不同的有针对性的营销策略，稳住高销量地区的客户，提高低销量地区自动售货机的知名度，增加人们的实用次数。

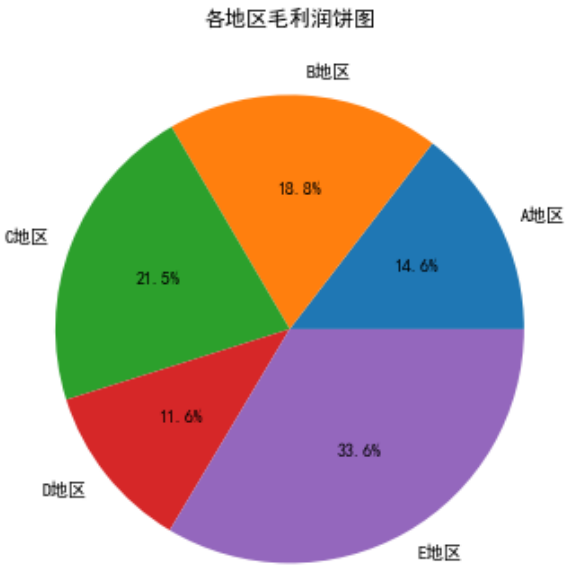

图 10 各地区毛利润占比

3.4 每月交易额均值气泡图

根据二级类和月份，完成商品分类，获取每月不同种类商品的交易额，并计算其均值。关联交易额均值与气泡大小，绘制气泡图如图 11。

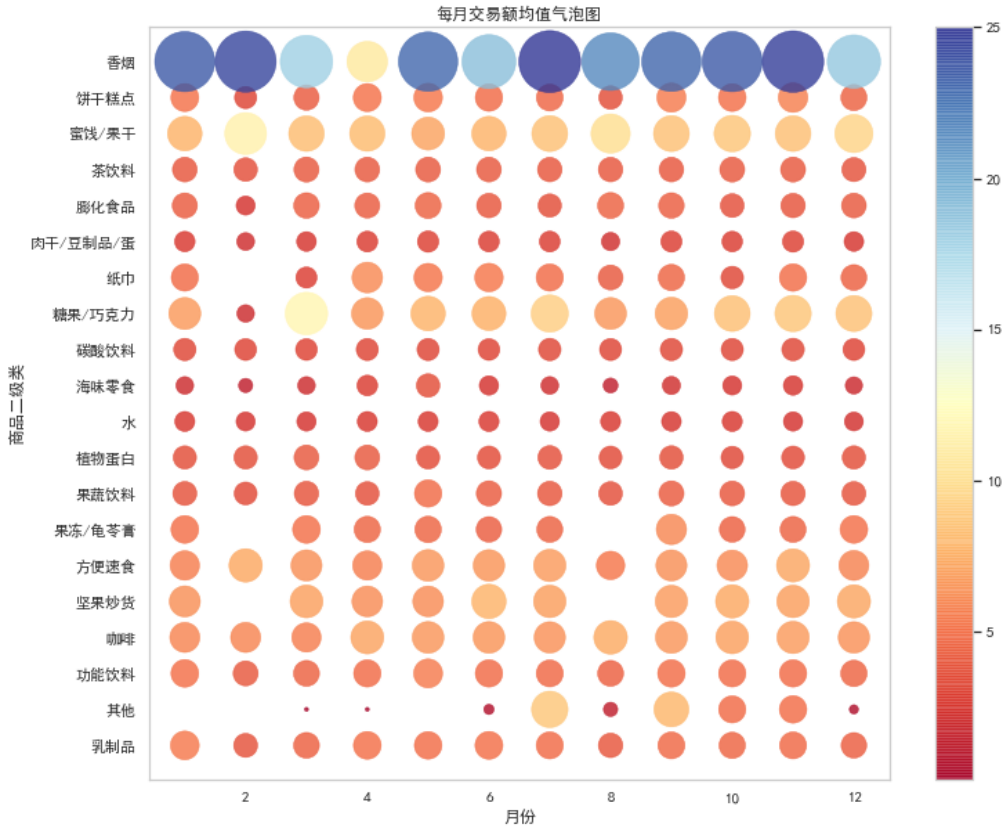

图 11 每月交易额均值气泡图

受价格影响，该图对人们对不同类型商品需求的反映程度较低。例如，由于香烟价格相比其他商品价格较高，香烟类商品的交易额显著高于其余类商品的交易额。

3.5 C 地区 6、7、8 月订单量热力图

首先，获取 C 地区自动售货机的订单数据，根据日期和小时数据分组，计算各组内的订单量。然后，生成以小时为横轴、日期为纵轴、订单量为值的矩阵。接着，根据矩阵生成热力图如图 12-14。

根据热力图，可大概推测出人们使用自动售卖机购买商品的时间段多数集中在下午和傍晚。

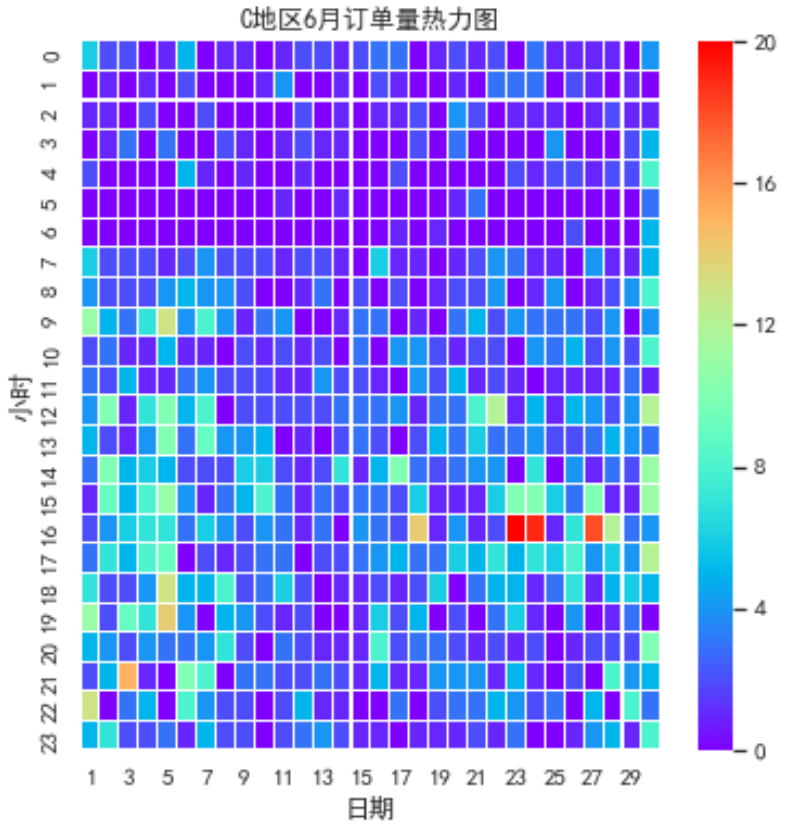

图 12 C 地区 6 月订单量热力图

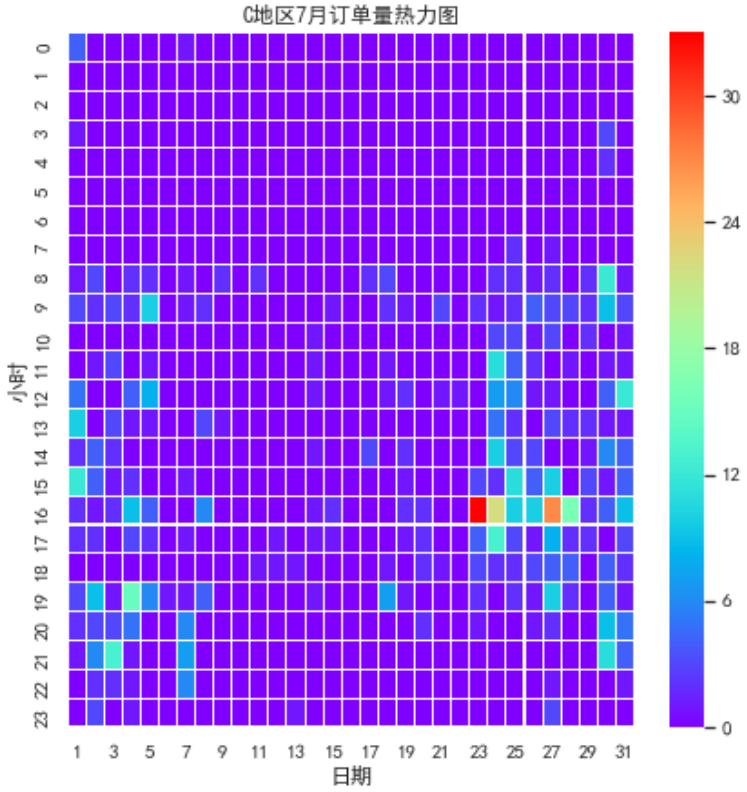

图 13 C 地区 7 月订单量热力图

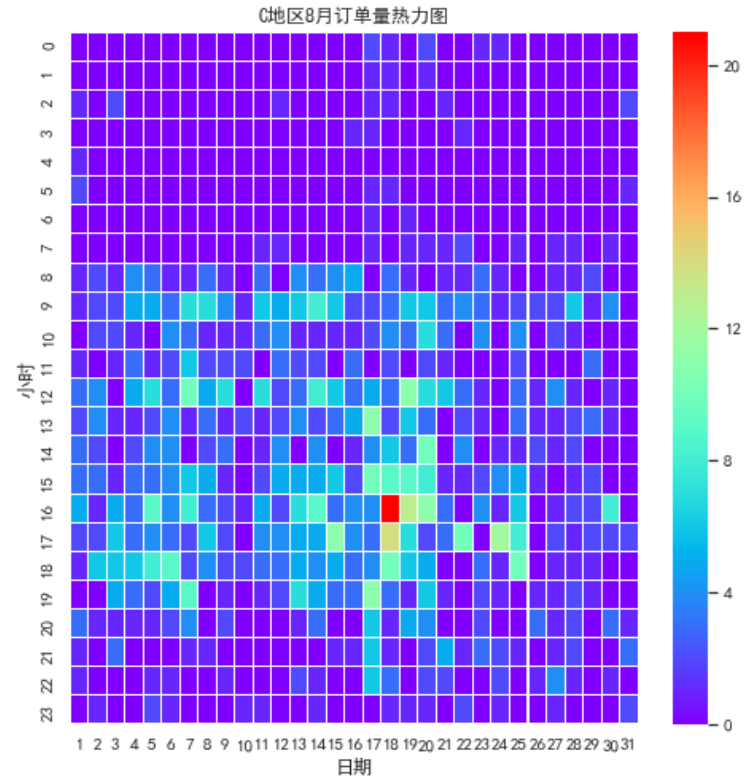

图 14 C 地区 8 月订单量热力图

四、生成自动售货机画像

4.1 生成饮料类商品标签

把售货机按地区及大类分类，得到每台自动售货机的饮料类商品数据，并获取商品销量的描述性统计。根据描述性统计，把销量在前 25%的商品归类为畅销商品，销量在后 25%的商品归类为滞销商品，销量排名在 25%-75%的商品则归类为正常商品。结果保存至”task3-1A.csv”、” task3-1B.csv”、…、” task3-1E.csv”，生成商品标签示例如表 3。

表 3　A 地区饮料类商品标签示例

序号	饮料类商品	订单量	标签
1	怡宝纯净水	531	畅销
2	东鹏特饮	448	畅销
3	雪碧 330ml	67	正常
4	500ml 统一阿萨姆奶茶	65	正常
5	可口可乐 330ml（罐）	5	滞销
6	脉动（青柠）	5	滞销
……			

4.2 生成自动售货机画像

通过在线词云工具，输入各地区商品销量数据，得到词云如图 15-19。

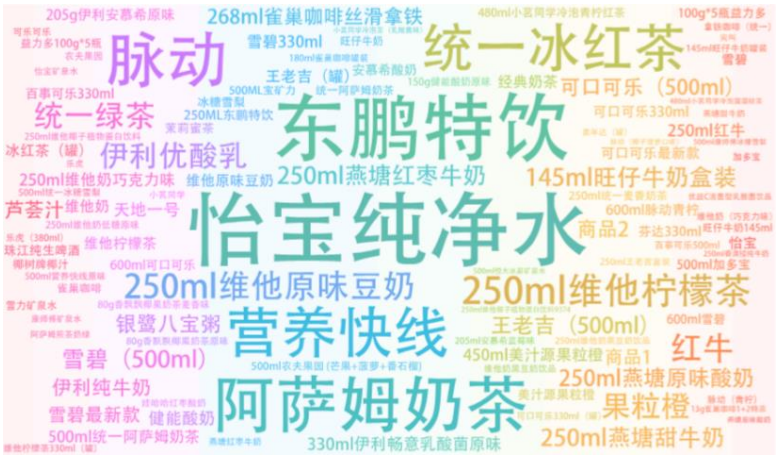

图 15 A 地区售货机销售商品词云

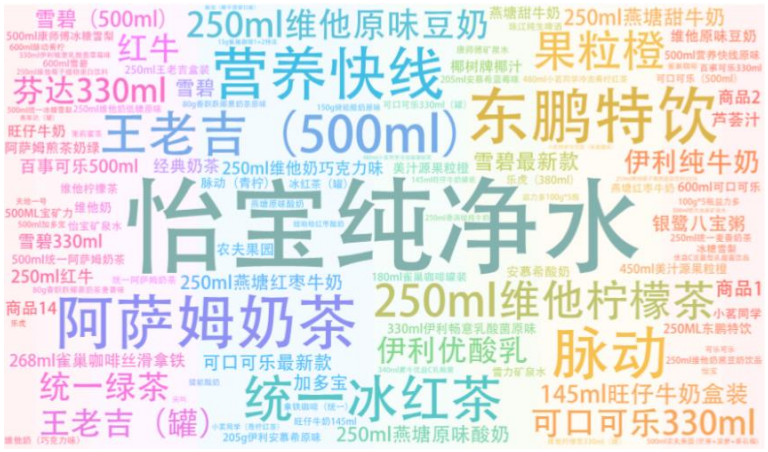

图 16 B 地区售货机销售商品词云

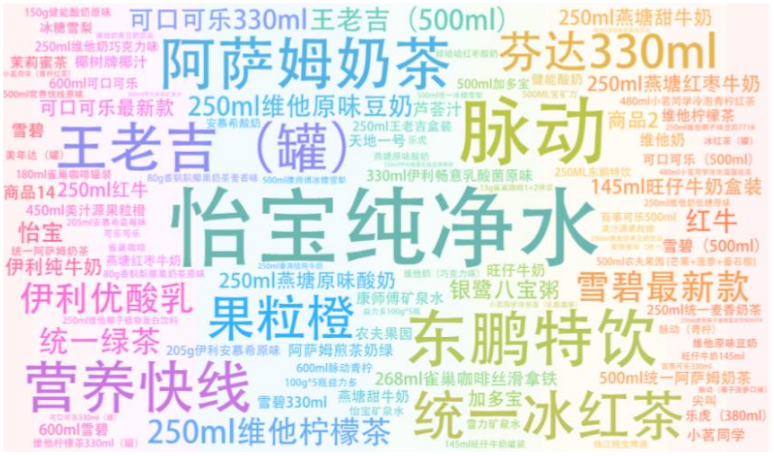

图 17 C 地区售货机销售商品词云

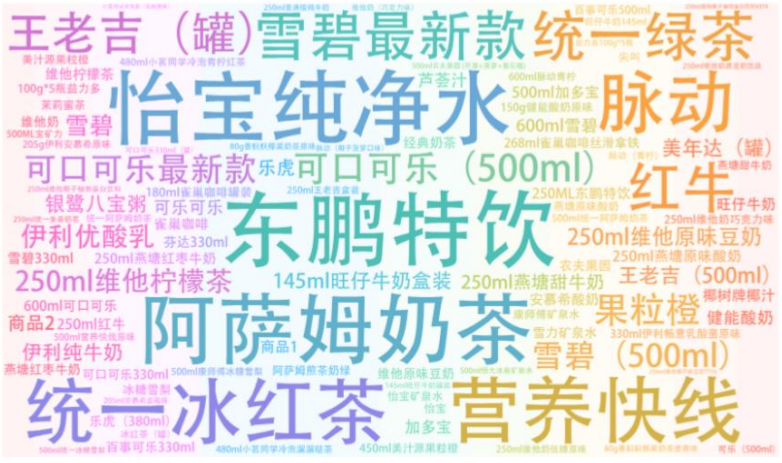

图 18 D 地区售货机销售商品词云

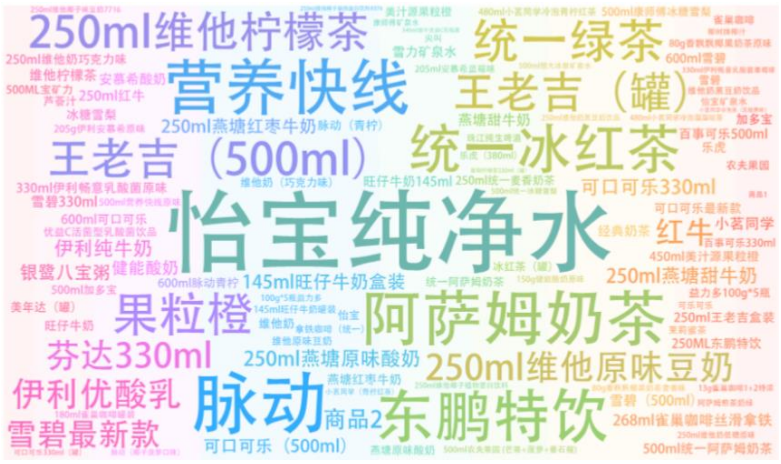

图 19 E 地区售货机销售商品词云

对比各地区商品销量词云图可见，各地区销量较高的几款饮料基本不变，没有显著差异。其中，怡宝纯净水在各地区都很受欢迎，其次就是奶茶、冰红茶、东鹏特饮等茶类饮料、乳制品和功能饮料。

为了进一步了解各地区商品销量结构，在 4.1 的数据集中加入二级类标签，通过二级类标签分组记录来观察各地区是否存在不同类型的饮料偏好，结果如表 4。

表 4 各地区商品销售数据

二级类	A地区销量	B地区销量	C地区销量	D地区销量	E地区销量
乳制品	1269	1660	1661	969	3030
其他	6	5	4	3	4
功能饮料	1109	1337	1783	1028	2519
咖啡	154	153	151	105	330
方便速食	103	120	166	83	220
果蔬饮料	318	461	546	236	942
植物蛋白	394	492	476	225	929
水	648	1408	1216	423	1859
碳酸饮料	693	1252	1276	939	2173
茶饮料	1430	2144	2530	1501	3959

通过二级类标签对销量进行统计可见，各地区的人群各地区销量最高的二级类饮料商品皆为茶饮料，其次是乳制品、功能饮料及水。对比词云可发现，怡宝纯净水虽然销量最高，但水并非占据销量第一的商品二级类。因此可以猜测，自动售货机售卖的水的品牌种类相对较少，而茶饮料等商品的种类繁多，使人们在品牌和口味上有更多的选择。于是出现了单一品牌的茶饮料销量不如单一品牌的水，但总体茶饮料销量却远高于水销量的现象。

结合各地区商品销售数据和词云，向商家提出以下建议：

- 可适当加大茶饮料、乳制品、功能饮料及水等热销商品的进货量，对于需求量较小的商品，可适量下调存货量，合理调整运输和储存空间；
- 可在需求量高地区（如 E 地区）增加自动售货机数量，加大自动售货机的放置密度，使人们消费更加便利；
- 对于销量较低的地区，可以实地考察售货机附近的人流量等因素，以期排除因选址不佳而导致销量不高的问题，除此以外也可适当进行促销活动提高知名度和刺激人们的消费；
- 根据地区属性的不同，人群对不同类型，甚至是不同包装方式的饮料有不同的需求，商家在部署自动售货机时应把选址附近的人群类型作为参考变量。

五、业务预测

5.1 预测原理

预测是根据事物发展过程的历史和现实，综合各方面的信息，运用定性和定量的科学分析方法，揭示事物发展的客观规律，指出其可能的发展途径及可能的发展结果。预测就是根据系统或类似系统过去和现在已经发生的状况，分析其发展和变化的规律并利用这个规律预计和描述系统将来某时期的状态或趋势。就是根据过去和现在来预计（估计）未来，根据已有的信息来推测未来的情况。

预测遵循以下基本原则：

一、连贯性原则：连贯性原则亦称惯性原则。所谓连惯性原则，就是从时间上考察事物的发展，其各个阶段具有连续性。

二、类推性原则：所谓类推性原则，就是根据过程的结构和变化所具有的模式和规律，可以推测出将来发展变化情况。

三、相关性原则：各种事物之间存在着直接或间接的联系，因此存在着相互影响、相互制约、相互促进的关系。

四、实事求是原则：准确可靠的调查统计资料和信息，是预测的依据。预测所依据的资料必须是准确可靠的，预测结果才能切合实际。

预测模型分为定性预测方法、时间序列分析、因果方法预测三类。时间序列分析是根据系统对象随时间变化的历史资料，只考虑系统变量随时间的变化规律，对系统未来的表现时间进行定量预测的方法，主要包括移动平均法、指数平滑法、趋势外推法等，该方法适于利用简单统计数据预测研究对象随时间变化的趋势等。

本项目中，将使用常见的进行时间序列预测的模型之一——ARIMA 模型，来根据附件提供的数据对每台售货机的每个大类商品在 2018 年 1 月的交易额进行预测。

5.2 ARIMA 模型

ARIMA 模型全称为自回归移动平均模型(Autoregressive Integrated Moving Average Model，简记 ARIMA)，是由博克思·詹金斯于 70 年代初提出的一著名时间序列预测方法，所以又称为 box--jenkins 模型、博克思·詹金斯法。其中，ARIMA(*p,d,q*)称为差分自回归移动平均模型，AR 是自回归，P 为自回归项。ARIMA 模型可分为 3 种：自回归模型（简称 AR 模型）、滑动平均（模型简称 MA 模型）和自回归滑动平均混合模型（简称 ARIMA 模型）。

以时间序列的自相关分析为基础，ARIMA 模型在经济预测过程中既考虑了经济现象在时间序列上的依存性，又考虑了随机波动的干扰性，对于经济运行短期趋势的预测准确率较高，是应用比较广泛的方法之一。建模基本步骤如下：

1. 获取被观测系统时间序列数据；
2. 对数据绘图，观测是否为平稳时间序列，对于非平稳时间序列要先进行 *d* 阶差分运算，化为平稳时间序列；
3. 经过第二步处理，已经得到平稳时间序列。要对平稳时间序列分别求得其自相关系数 ACF 和偏自相关系数 PACF，通过对自相关图和偏自相关图的分析，得到最佳的阶层 *p* 和阶数 *q*；
4. 由以上得到的 *d*、*q*、*pd*、*q*、*p*，得到 ARIMA 模型。然后开始对得到的模型进行模型检验。

5.3 应用 ARIMA 模型进行预测

售货机每月交易数据可以看作是随着时间的推移而形成的一个随机时间序列，通过对该时间序列上交易额的随机性、平稳性以及季节性等因素的分析，将这些单月交易额之间所具有的相关性或依存关系用数学模型描述出来，从而达到利用过去及现在的交易额信息来预测未来交易额情况的目的。

以 A 地区自动售货机为例，2017 年饮料类商品每月总交易额可视化处理如图 20。

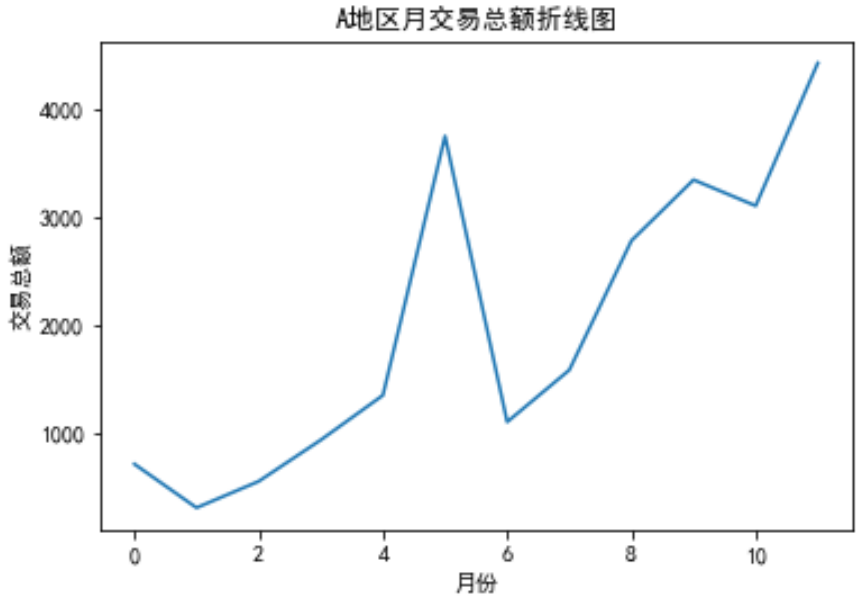

图 20 A 地区饮料类商品月交易额

5.3.1 对序列取对数和作差分处理，形成稳定随机序列

ARIMA 模型建模的基本条件是要求待预测的数列满足平稳的条件，即个体值要围绕序列均值上下波动，不能有明显的上升或下降趋势，如果出现上升或下降趋势，需要对原始序列进行差分平稳化处理。

由图 21 可见，一阶差分的时间序列的均值和方差已经基本平稳，二阶差分后的时间序列与一阶差分相差不大，并且二者随着时间推移，时间序列的均值和方差保持不变。因此可以将差分次数 *d* 设置为 1。

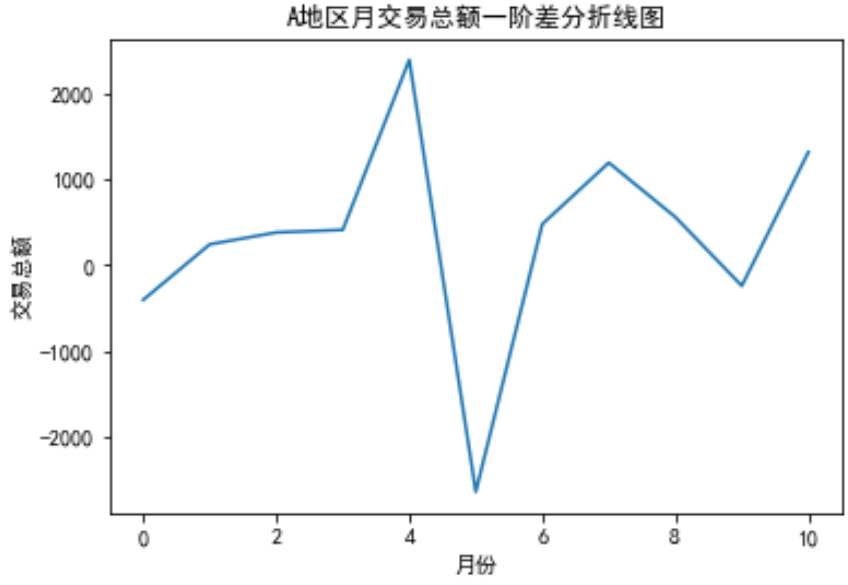

图 21 A 地区饮料类商品月交易额一阶差分

5.3.2 模型参数的估计

时间序列预测模块的自相关分析包括对自相关系数和偏相关系数的分析，通过对比分析从而实现对时间序列特性的识别，如图 22。从结果可知，自相关函数一步截尾，偏自相关函数为拖尾，自相关函数通过白噪声检验。根据变换数列的自相关函数和偏自相关函数的特点，并经过反复测试，对 ARIMA 模型的参数进行估计，层数 p 设为 0，阶数 q 设为 1，则构建模型 ARIMA(0,1,1)。

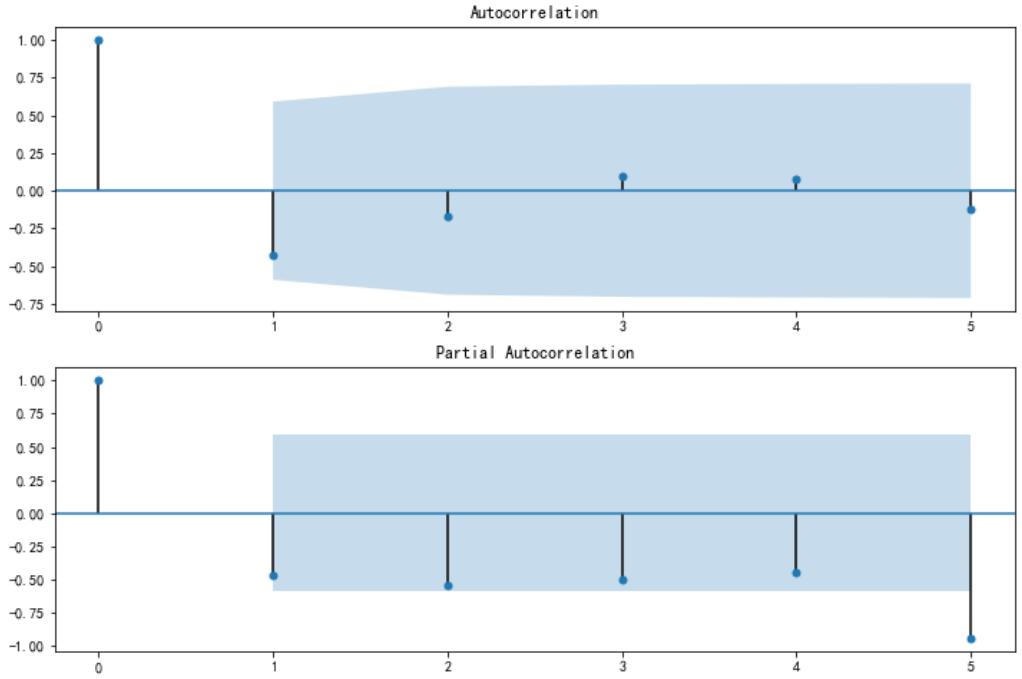

图 22 自相关图与偏自相关图

接下来，对模型进行 D-W 检验。德宾-沃森检验,简称 D-W 检验，是目前检验自相关性最常用的方法，但它只使用于检验一阶自相关性。当 DW 值显著的接近于 0 或 4 时，则存在自相关性，而接近于 2 时，则不存在（一阶）自相关性。检验结果是 2.2250，说明不存在自相关性，可认为模型合理，可使用 ARMA(0,1,1)对数据进行预测。

5.3.2 应用模型预测

利用上面确定的模型进行预测，如图 23，蓝色折线为原始序列，红色折线为预测序列。预测模型 2017 年饮料类商品 12 月交易额的拟合值是 3751.66 元，跟实际交易额 4424.6 元比较，误差为-15.2%，表明预测模型拟合度不高，预测模型应用价值有限。

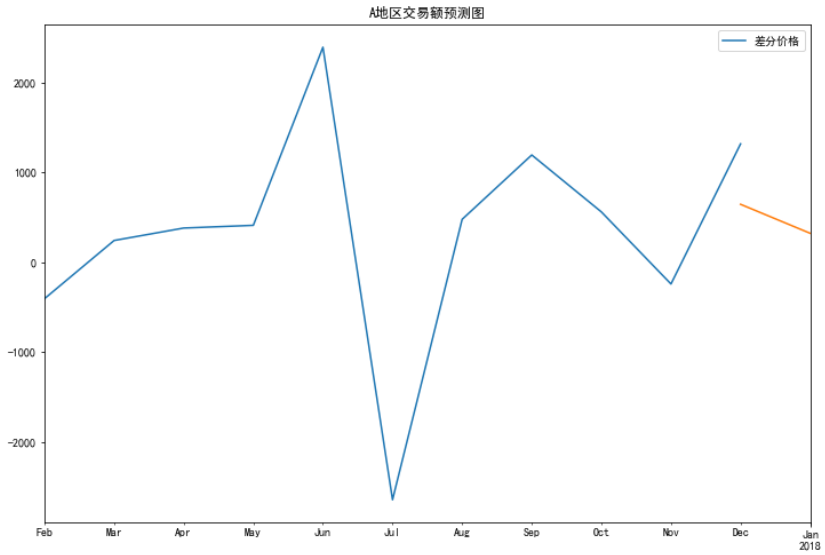

图 23 A 地区饮料类商品差分序列预测图

预测得到 2018 年 1 月份的差分价格后，根据

$$M_i = M_{i-1} + D_i$$

计算 18 年 1 月交易总额，其中 M_i 为月度交易总额，D_i 为当月差分价格。计算各地区的大类商品预测值并统计得到表 5 。

表 5 各地区大类商品预测交易额

地区（大类）	2017 年 12 实际交易额	2017 年 12 月预测交易额	201 年 1 月预测交易额
A（饮料）	4424.8	3751.62	4748.87
A（非饮料）	3162.3	2687.5	3370.89
B（饮料）	5845.5	6018.62	6169.57
B（非饮料）	2258.6	2921.04	2501.55
C（饮料）	6608.1	6112.72	6932.17
C（非饮料）	2776.4	3281.81	3031.29
D（饮料）	3967.4	3450.71	4291.47
D（非饮料）	1973.8	1705.23	2011.21

E（饮料）	9316.3	10423.9	14383.46
E（非饮料）	4241.2	5145.16	4719.53

由表 5 可见，2017 年 12 月的实际交易额与预测交易额之间存在着一定的误差，且随着真实交易额的增加，预测交易额的误差也会增大。2018 年 1 月的预测交易额较上一年 12 月均有不同程度的提升，且维持了各地区原本的相对水平，即热销地区的预测值比相对滞销地区的预测值更高，因此模型的预测结果具有一定的合理性。但注意到，自动售货机交易额可能存在以季度、年为周期的特征，这是模型在构建过程中无法获取的，因此预测数据存在着一部分非系统误差。总体来说，模型的预测结果是高于 2018 年 1 月交易总额的期望，即模型是一个乐观模型，实际上的交易额数据应该会比预测值低。

5.3.3 误差分析

导致误差较大的一个重要原因是数据量不足，原始数据中仅有 2017 年一年的数据，且部分数据存在大量的缺失，模型无法提取数据以季度、年为周期的特征，这种信息的损失使得模型的预测效果大打折扣。若需要更精准的预测模型，则需对数据的时间长度进行拓展，例如获取历史数据或在未来持续关注并记录数据，以此不断调整模型提高精度。