新零售-无人售货机商务数据分析

孙建存

目录

1 数据预处理与分析

1.1 数据预处理

1.1.1 去除多余属性

1. “订单号”、“状态”和“提现”属性不包含有效信息，可以删除。
2. 经检测，“应付金额”和“实际金额”全部相等，所以删除“应付金额”列，修改“实际金额”属性名为“金额”。
3. 同一个地区，只包含一个“设备 ID”，为方便计算，删除“设备 ID”列。

	金额	商品	支付时间	地点
0	4.5	68g好丽友巧克力派2枚	2017/1/1 00:53	D
1	3.0	40g双汇玉米热狗肠	2017/1/1 01:33	A
2	5.5	430g泰奇八宝粥	2017/1/1 08:45	E
3	5.0	48g好丽友营愿香烤原味	2017/1/1 09:05	C
4	3.0	600ml可口可乐	2017/1/1 09:41	B

图 1-1-1 去除多余属性后的部分数据

1.1.2 检测异常值

将销售数据按“商品”属性进行分组，使用箱线图检测商品金额中异常的数据，获取异常数据的索引，并全部去除。

1.1.3 去除金额为 0 的数据

经检测发现销售数据在去除异常值后，仍包含部分商品金额为 0，所以去除该部分数据。

去除的异常值数据和金额为 0 的数据共 7004 行。

1.1.4 添加商品信息

为方便后续统计分析，将附件 2 中的商品信息添加到附件 1 中，属性名分别修改为“商品大类”和“商品二级类”。修改后的数据部分如图 1-1-2 所示。

	金额	商品	支付时间	地点	商品大类	商品二级类
0	4.5	68g好丽友巧克力派2枚	2017/1/1 00:53	D	非饮料	饼干糕点
1	3.0	40g双汇玉米热狗肠	2017/1/1 01:33	A	非饮料	肉干/豆制品/蛋
3	5.0	48g好丽友营愿香烤原味	2017/1/1 09:05	C	非饮料	膨化食品

图 1-1-2 添加商品信息后的部分数据

1.1.5 按地点提取数据

按照地点 A,B,C,D,E 提取每台售货机对应的销售数据，并保存在 csv 文件中，文件名分别为”task1-1A.csv”、 ”task1-1B.csv”、…、 ”task1-1E.csv”

1.1.6 其它处理

对附件 1 中的数据观察可发现不存在缺失值，因此无需进行缺失值处理，但经后续处理发现“支付时间”列存在数据错误，且错误数据只有 1 个，因此将原来的“2017/2/29”直接修改为“2017/3/1”。

1.2 五月份销售信息

提取各个地点的 5 月份的销售数据，计算各个地点的售货机的订单量和交易额。然后根据所有数据计算 5 月份总的订单量和交易额。计算结果如表 1-2-1.

表 1-2-1 售货机 5 月份的销售情况

指标 \ 地点	A	B	C	D	E	总和
交易额	2734.1	2902.5	2817.3	1872.2	4231.8	14557.9
订单量	679	734	665	499	1048	3625

由上表可以看出，地区 E 的售货机较其它地区的售货机销售情况明显较好，而 D 区售货机销售情况最差，A,B,C 区销售情况相似。

1.3 平均交易额和日均订单量

首先在每个地区的销售数据中按月份拆分数据，统计每台售货机每月的总交易额和总订单量，两者做除法运算可得每单的平均交易额。在计算日均订单量时发现个别月份的有效数据过少，因此若按照每个月包含的天数计算得出的数据不太合理，所以应该实现统计每个月份中有货物出售的有效天数，再与总订单量做除法运算可得日均订单量。

表 1-3-1 每台售货机每月的每单平均交易额

时间 \ 地点	A	B	C	D	E
一月	3.951	3.546	3.916	3.545	3.926
二月	3.741	3.26	3.8	3.08	3.624
三月	3.379	3.53	3.514	3.9	4.193
四月	3.815	3.62	3.946	3.509	3.841

五月	4.027	3.954	4.237	3.752	4.038
六月	3.68	3.831	3.942	3.726	3.605
七月	3.716	4.031	3.793	3.8	3.743
八月	3.24	3.547	3.869	3.29	3.68
九月	3.946	3.909	4.118	3.626	3.851
十月	3.706	3.814	3.901	3.565	3.607
十一月	4.089	3.952	4.092	3.64	4.013
十二月	3.55	3.579	3.766	3.428	3.951

表 1-3-2 每台售货机每月的日均订单量

地点 时间	A	B	C	D	E
一月	12	15	12	11	11
二月	4	7	7	5	8
三月	47	40	40	44	45
四月	13	17	21	13	25
五月	22	24	22	16	34
六月	50	54	53	31	76
七月	16	11	24	10	23
八月	20	30	38	22	52
九月	30	53	50	30	121
十月	45	58	63	35	81
十一月	34	63	59	38	148
十二月	59	66	70	50	91

由表 1-3-1 可知，每台售货机在不同月份的平均交易额大致相似，且随时间变化无明显波动。

由表 1-3-2 可知，在大部分月份中，E 地区的日均订单量高于其它地区，而 D 区订单量相对较少。每台售货机随时间变化大致呈递增趋势，但七、八月份整体销量不佳。

2 数据分析与可视化

2.1 六月销量前 5 的商品销量

根据要求，首先将“支付时间”属性转换为时间格式，然后提取 6 月份的商

品销售数据，根据“商品”属性统计所有商品的销量，按降序排列数据，提取前五组数据即可。

所绘制柱状图如图 2-1-1.

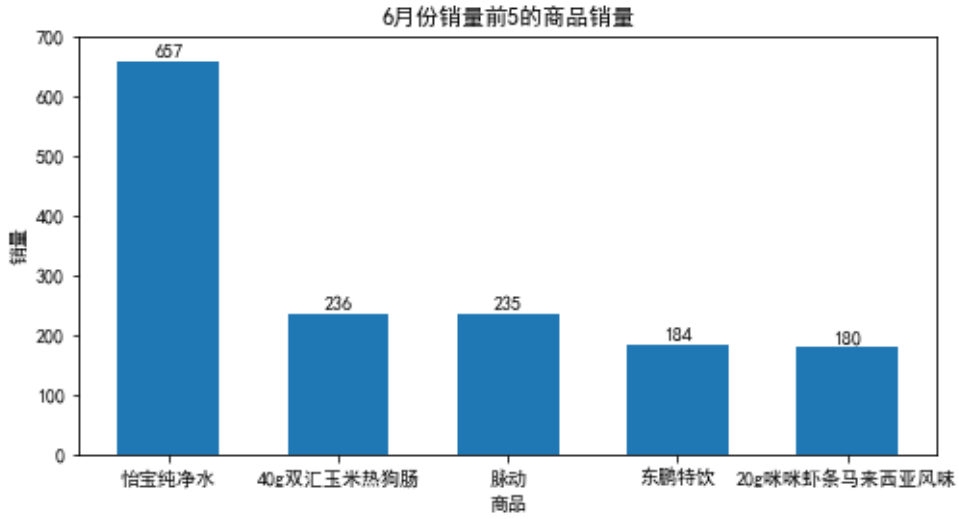


图 2-1-1 六月份销量前 5 的商品及销量

由上图可以看出 6 月份销量前五的商品分别为怡宝纯净水、40g 双汇玉米热狗肠、脉动、东鹏特饮和 20g 咪咪虾条马来西亚风味。其中怡宝纯净水的销量非常好，是排名第二的商品销量的三倍左右。40g 双汇玉米热狗肠与脉动销量相似，东鹏特饮和 20g 咪咪虾条马来西亚风味销量相似。

2.2 总交易额折线图和月环比增长率柱状图

2.2.1 总交易额折线图

实际上在任务 1.2 中已经计算了每台售货机每月的总交易额，并以字典形式进行了存储，所以只需获取相应的索引即可提取数据。而索引是每个地区的名称和各自对应的月份，按照索引提取信息，并将信息保存在列表中，然后依次绘制折线图，结果如图 2-2-1 所示。

由折线图可以看出，每台售货机在 1-2 月份交易额变化趋势相同，且数额相似，在 3-12 月份 E 地区每月的总销售额明显大于其它地区，且与其他地区交易差额逐渐增大，但是 E 地区总交易额增长的间隙伴随有较大的回落，交易额变化不稳定。其它地区发展趋势大致相同，其中 D 地区销售情况略佳。从整体上看，每个地区的售货机随时间推移呈增长趋势，但是在七月份都有幅度较大的回落，商家应该在七月份减少供货量，在其它月份逐渐增大供货量。

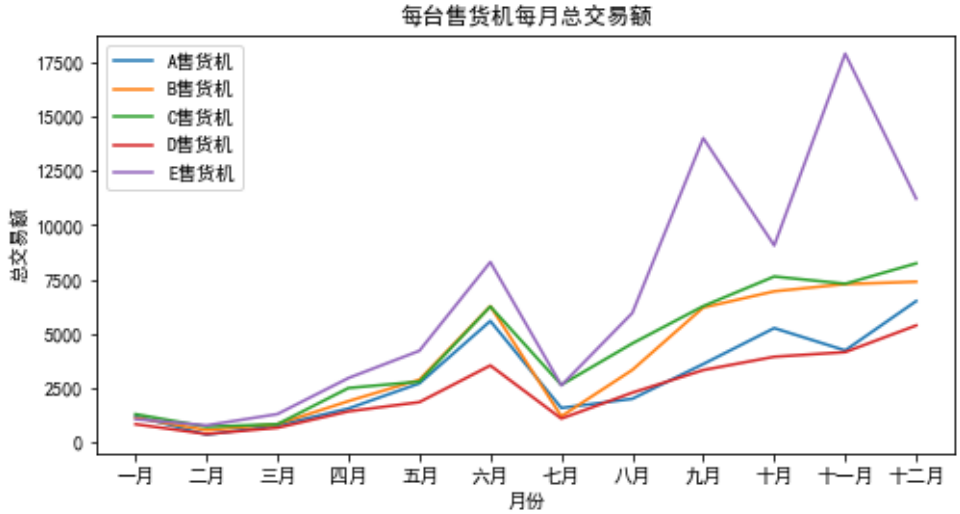
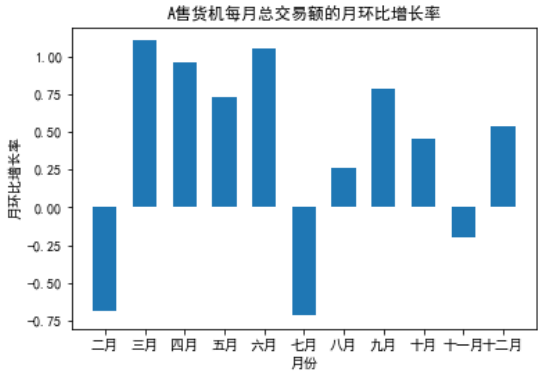
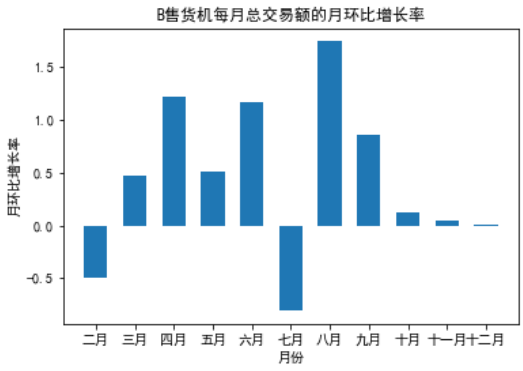


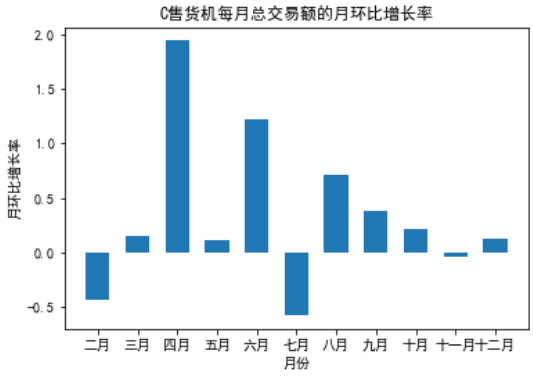
图 2-2-1 每台售货机每月总交易额折线图

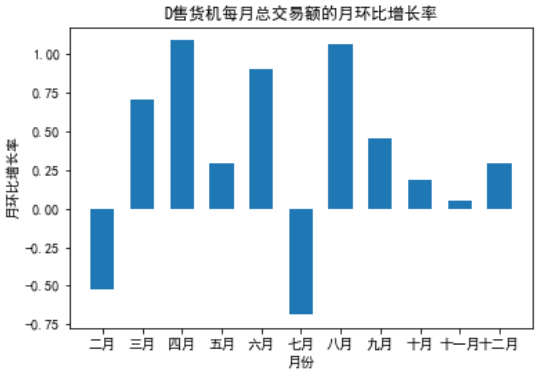
2.2.2 交易额月环比增长率

根据已经获取的每台售货机每月的总交易额可以很方便地计算交易额月环比增长率。









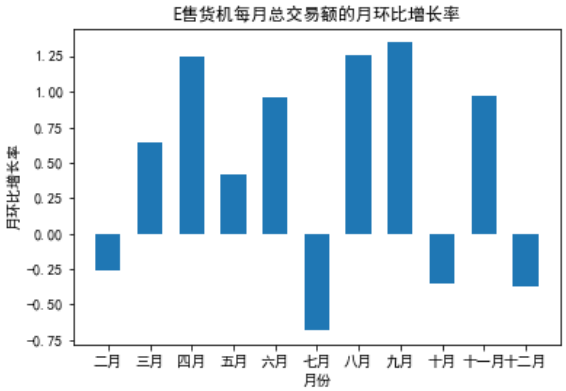


图 2-2-2 每台售货机每月总交易额月环比增长率柱状图

从五张柱状图可以看出，每个地区的售货机在 2 月份和 7 月份都呈现负增长，且 7 月份负增长尤为突出。除此之外，E 地区的售货机总交易额在 10 月份和 12 月份也呈现负增长。A 地区和 C 地区的售货机的月环比增长率大致呈下降趋势。从数值上看，C 地区除 4 月和 6 月增长幅度较大外，其它月份变化不明显，销售情况较为稳定，其他地区变化较大。

2.3 毛利润饼状图

首先计算总毛利润，将读取的数据按照商品大类分组，并计算各自的总交易额，按照饮料类商品毛利率 25%和非饮料类商品毛利率 20%计算各自的毛利润，然后求和可以得到总毛利润。按照这种方法可以计算出各个地区饮料类和非饮料类商品的毛利润。然后做除法运算得出各个地区的毛利润所占比例。

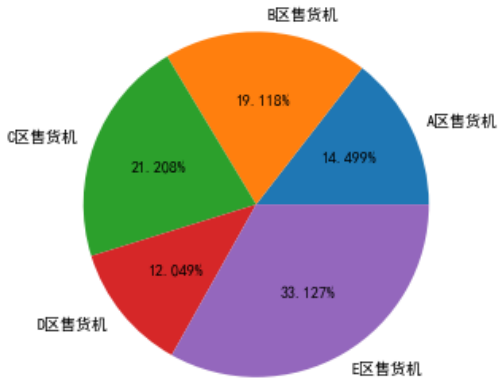


图 2-3-1 每个地区的毛利润比例

由饼状图可以看出，E 地区毛利润所占比例最大，占比 33.127%，B、C 地区售货机占比相似，在 20%左右，是五个地区的平均值，A、D 地区占比最少，毛

利润低于平均值。

2.4 每月交易额均值气泡图

要绘制气泡图，所需数据集的属性包括“支付时间”、“金额”和“商品二级类”，按照商品二级类和月份对数据进行分组，计算每组的总交易额。然后将总交易额与每月的天数做除法运算，可以得到每类商品的每月交易额均值。以月份为横轴，二级类为纵轴绘制气泡图。

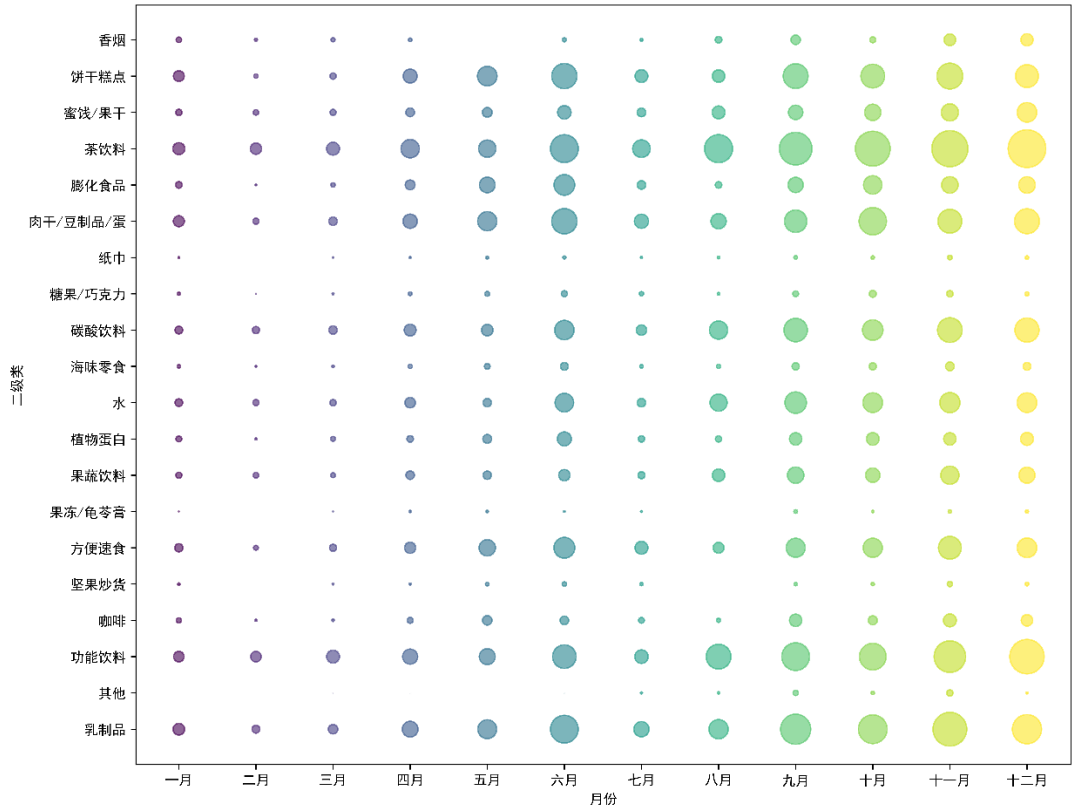


图 2-4-1 每月交易额均值气泡图

由气泡图可知，茶饮料、功能饮料和乳制品在每个月的交易额均值排列中非常突出，其次饼干糕点、肉干/豆制品/蛋和方便速食的交易额均值也比较明显。而“香烟”、“纸巾”、“糖果/巧克力”、“海味零食”、“果冻/龟苓膏”、“坚果炒货”和“其他”商品保持在一个较低的水平。

2.5 六到八月份的订单量热力图

提取 C 地区售货机的销售数据，按支付时间的日期和小时对数据进行分组，每组包含的数据个数就是每组的订单量。

以日期为横轴，以小时为纵轴，绘制 6-8 月份的热力图。

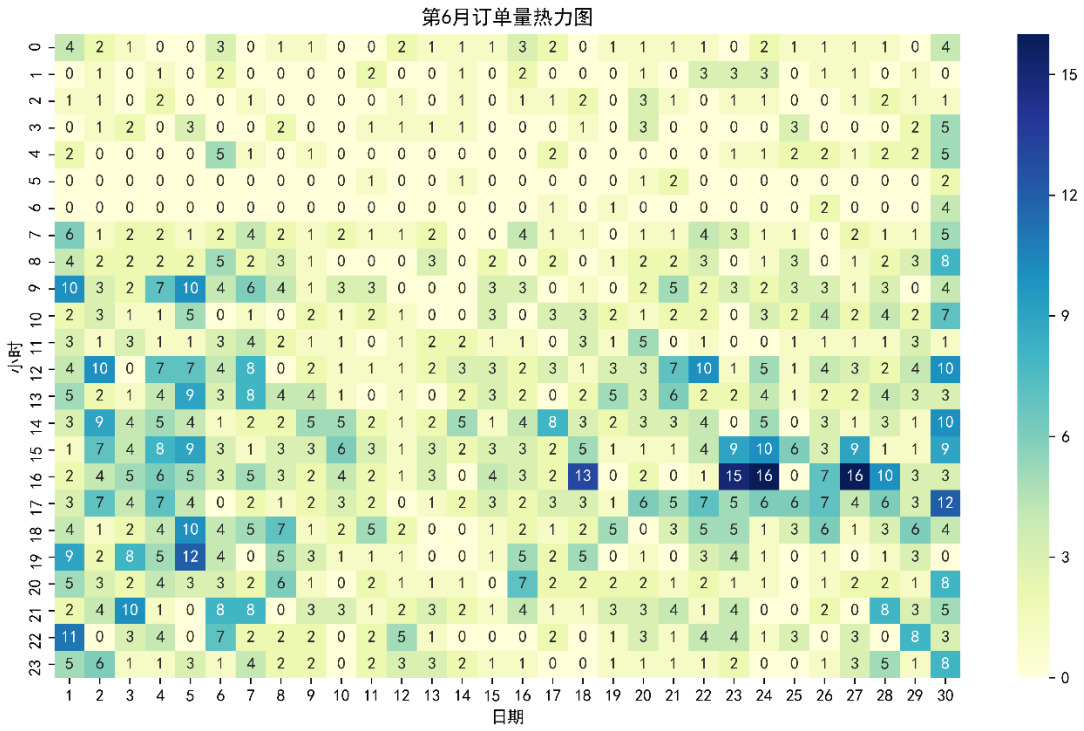


图 2-5-1 C 地区售货机 6 月份订单量热力图

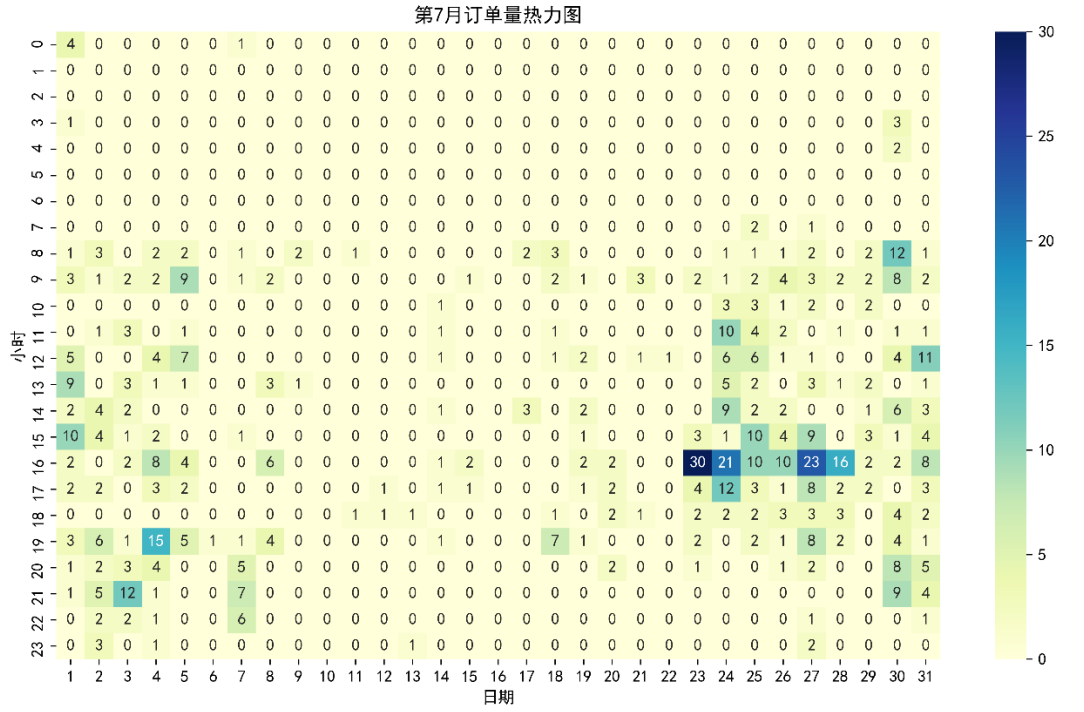


图 2-5-2 C 地区售货机 7 月份订单量热力图

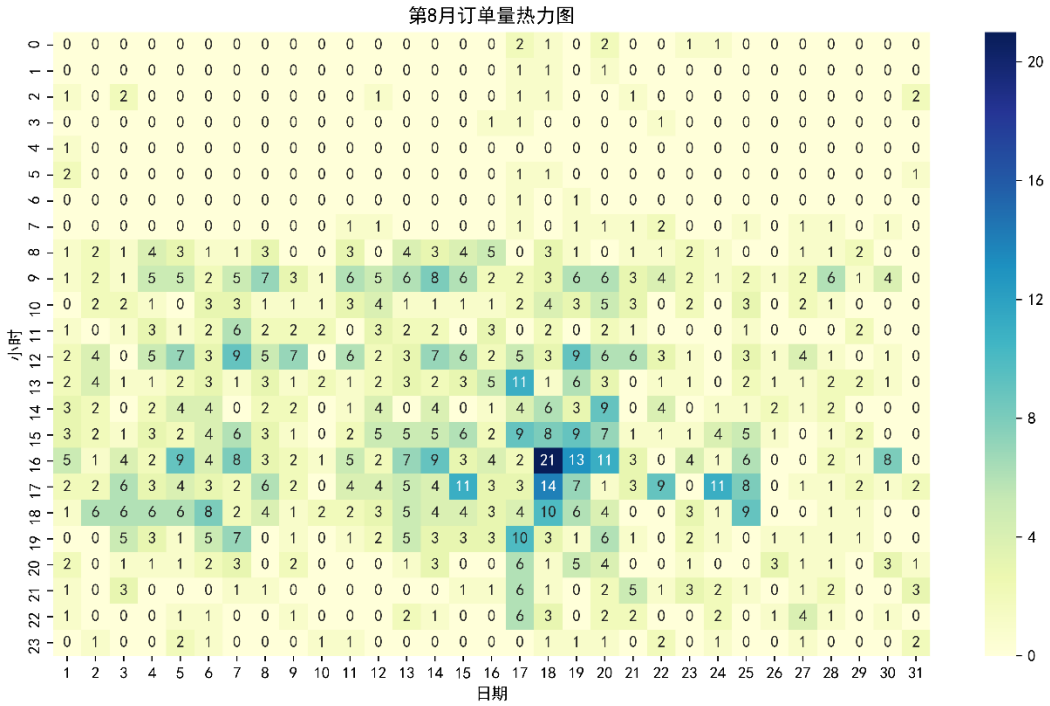


图 2-5-3 C 地区售货机 8 月份订单量热力图

从日期上看，6 月份和 7 月份的订单量集中在上旬和下旬，而 8 月份中旬的订单较为集中。

从小时上看，6-8 月份大都集中在下午 4 时左右。另外，6 月份上旬在晚上的订单也较多。商家应该注意在这些高峰期前进行补货。

从整体上看，7、8 月份是一个淡季，订单量偏少。这点在折线图上也可以看出。

3 自动售货机画像

3.1 商品标签

给出每台售货机所有商品（饮料类和非饮料类）的商品标签

在对商品贴上标签之前，首先对销量数据进行分析。提取所有商品的销量数据，商品销量最低为 1，最高为 4964，将此区间划分为 10 段，对每段上的数据做计数统计，绘制柱状图如 3-1-1 所示。

由柱状图可以看出，绝大部分商品的销量在 1 到 500 之间，少部分商品的销量超过了 500，所以将 500 定为上限，即商品总销量大于 500 可认为是热销，在实际计算中，将 500 这个标准平均分配到 5 个地区的售货机，即每个地区的商品销量大于 100 可认为是热销。总销量的下限定在 1 到 500 之间的中位数，即商品

总销量低于该值可认为是滞销，大于该值且小于 500 的可认为是正常。

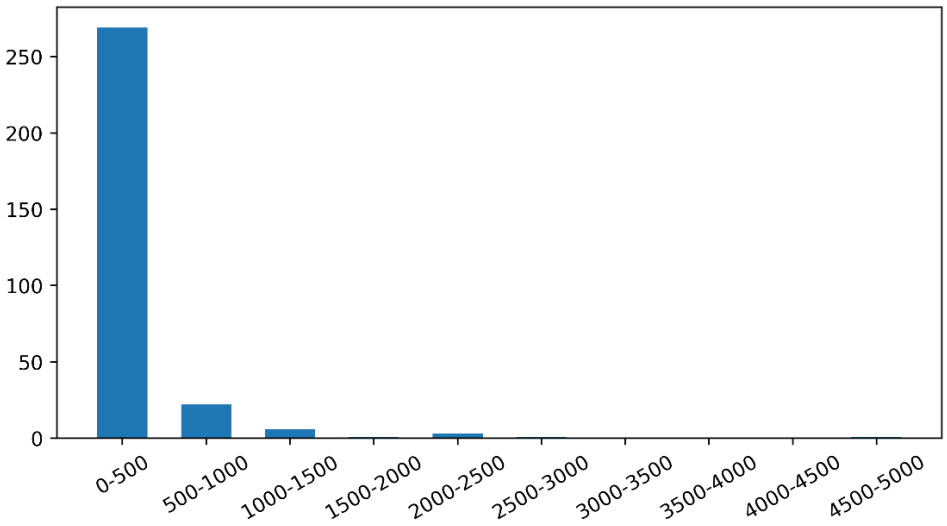


图 3-1-1 销量分布情况

以上述标准对每个地区的每种商品添加销售情况标签。结果如图 3-1-2 所示。（详细结果见附件）

	商品	销量	标签
0	100g*5瓶益力多	53	正常
1	100g卫龙点心面黑椒牛排味	9	滞销
2	100g果王咸柑桔罐装	14	正常
3	103g康师傅红烧牛肉面	17	正常
4	107g出前一丁桶面酱香牛肉王	12	正常

图 3-1-2 部分标签数据

3.2 标签扩展

将附件 2 数据中的“二级类”属性做出修改，使其能对商品做出更精准的分类，如将商品“劲仔小鱼”的二级类由“肉干/豆制品/蛋”修改为“肉干”。

使用正则表达式去除“商品”属性中的字母、数字和其它无用字符。

分别将上述两种数据按商品对应添加，修改属性名为“标签 1”和“标签 2”。

	商品	销量	标签	标签1	标签2
0	100g*5瓶益力多	53	正常	乳制品	益力多
1	100g卫龙点心面黑椒牛排味	9	滞销	豆制品	卫龙点心面黑椒牛排味
2	100g果王咸柑桔罐装	14	正常	果干	果王咸柑桔罐装

图 3-2-1 标签扩展后的部分数据

3.3 绘制售货机画像

按照扩展后的“标签 1”和“标签 2”生成画像。（标签 1 数据合并后只包含 24 种，因此对标签 1 数据排序后取前 3 种代表热销种类商品。对标签 2 数据排序后取前 50 种商品）

图 3-3-1 A 地区售货机画像

图 3-3-2 B 地区售货机画像

图 3-3-3 C 地区售货机画像

图 3-3-4 D 地区售货机画像

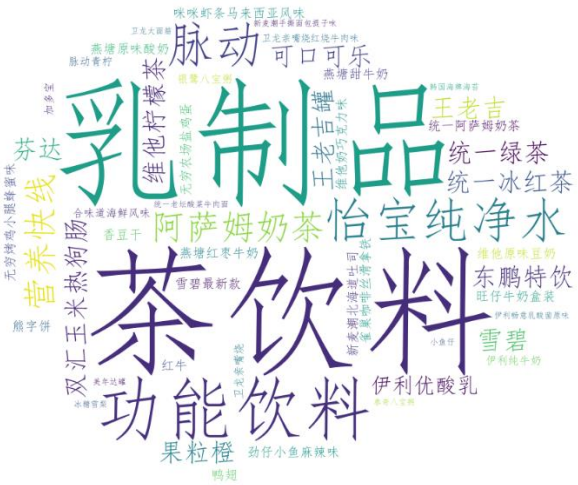


图 3-3-5 E 地区售货机画像

由售货机画像可以看出，“乳制品”、“茶饮料”和“功能性饮料”这三类商品在每个地区的销售排名中都比较靠前，占据了很重要的地位，因此商家需要注意这三类商品的供货。

就具体的商品而言，A 地区热销的几款商品为“怡宝纯净水”、“东鹏特饮”、“阿萨姆奶茶”、“脉动”、“营养快线”、“雪碧”和“统一冰红茶”；B 地区热销的几款商品为“怡宝纯净水”、“东鹏特饮”、“阿萨姆奶茶”、“脉动”、“营养快线”、“可口可乐”、“统一冰红茶”；C 地区热销的几款商品为“怡宝纯净水”、“脉动”、“阿萨姆奶茶”、“营养快线”、“王老吉罐”、“统一冰红茶”、“东鹏特饮”；D 地区热销的几款商品为“东鹏特饮”、“怡宝纯净水”、“阿萨姆奶茶”、“营养快线”、“统一冰红茶”、“可口可乐”、“脉动”；E 地区热销的几款商品为“怡宝纯净水”、“脉动”、“阿萨姆奶茶”、“营养快线”、“统一冰红茶”、“可口可乐”、“东鹏特饮”。可以看出每个地区最热销的几个商品大致相似，且全部为饮料产品。

3.4 营销意见

从日均订单量上看，五个地区在 1 到 7 月的日均销量基本相似，而 E 地区售货机在 8 到 12 月的销售量激增，尤其在 9 月和 11 月更是突破了日均 100 单，远远高于其他地区的售货机销量，因此商家在这个时期应该着重关注 E 地区售货机，需要及时检查货品销售情况，并进行补货。

从每月总交易额折线图上看，每个地区的售货机每月交易额大致呈递增趋势，但是七月份有一个大的回落，商家需要进行具体原因分析，并实施相应措施，如缩减供货量以减少运营成本，亦或者举行促销活动以增加收入。

根据商品的标签信息，商家应该根据商品的热销或 滞销情况，添加或减少相应的商品。

4 业务预测

自动售货机的经营者所给数据是 2017 年五个地区售货机的销售信息，要求预测 2018 年 1 月份的交易额。从现有情况看，商家所提供数据偏少，对一些问题不能确定，如商品的销售趋势是否以年为周期。如果以年为周期，则 1 月份会经历一次销售额下跌。因此如果商家提供更多年份的销售数据，则会增加预测数据的准确性。

假设商品的销售趋势不具有周期性。而销售信息是一个具有时间序列特性的数据，所以可以使用 ARMA 模型来进行预测。

4.1 ARMA 模型原理

ARMA 模型是研究时间序列的重要方法，由 AR 模型与 MA 模型混合而成。总的来说，AR 模型（自回归模型）是通过分析研究历史数据对当前数据的影响进行建模。MA 模型（移动平均模型）是用过去各个时期的随机干扰或预测误差的线性组合来得到当前预测值。

4.1.1 AR 模型

如果某个时间序列的任意数值可以表示成下面的回归方程，那么该时间序列服从 p 阶的自回归过程，可以表示为 AR(p)：

$$x_t = \phi x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + u_t$$

可以发现，AR 模型利用前期数值与后期数值的相关关系（自相关），建立包含前期数值和后期数值的回归方程，达到预测的目的，因此成为自回归过程。

4.1.2 MA 模型

如果某个时间序列的任意数值可以表示成下面的回归方程，那么该时间序列服从 q 阶的移动平均过程，可以表示为 MA(q)：

$$x_t = u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \phi_q u_{t-q}$$

可以发现，某个时间点的指标数值等于白噪声序列的加权和，如果回归方程中，白噪声只有两项，那么该移动平均过程为 2 阶移动平均过程 MA(2)。比较自回归过程和移动平均过程可知，移动平均过程其实可以作为自回归过程的补充，解决自回归方差中白噪声的求解问题，两者的组合就成为自回归移动平均过程，称为 ARMA 模型。

4.1.3 ARMA 模型

自回归移动平均模型由两部分组成：自回归部分和移动平均部分，因此包含

两个阶数，可以表示为 ARMA(p,q)，p 是自回归阶数，q 为移动平均阶数，回归方程表示为：

$$x_t = \phi x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \phi_q u_{t-q}$$

从回归方程可知，自回归移动平均模型综合了 AR 和 MA 两个模型的优势，在 ARMA 模型中，因此，该模型更为有效和常用。

4.2 预测过程

一般来说，ARMA 模型的训练大致包括：ADF 检验、平稳化处理和模型预测。ADF 检验用于检测数据平稳性，由于 ARMA 模型对数据的平稳性要求很高，因此不平稳的数据无法得到较好的预测结果，如果检验输出的 p-value 较大，则表明数据不具有平稳性。如果数据不具有平稳性，需要先做平稳化处理，常用方法有差分法和对数法。平稳化处理之后再次进行 ADF 检验，检验通过后进行白噪声检验，若随机概率大于 0.05，则不用再进行序列分析，如果低于 0.05，则可以将序列数据用于预测。

以 A 地区售货机为例，首先按照商品大类分为饮料类商品和非饮料类商品，再依据月份和日期进行销售额统计，得出 2017 年每天的两类商品的销售额。

对饮料类商品和非饮料类商品分别进行分析。平稳性检验中，饮料类商品和非饮料类商品的 adf 值均小于 1%临界值，且 p 值远小于 0.05，因此认为数据具有平稳性。白噪声检验结果中随机概率小于 0.05，因此可认为是非白噪声序列。

寻找最优参数，发现两类商品的 p 和 q 都为 1 时模型最佳，使用该模型进行预测，预测结果如图 4-2-1、图 4-2-2 所示.

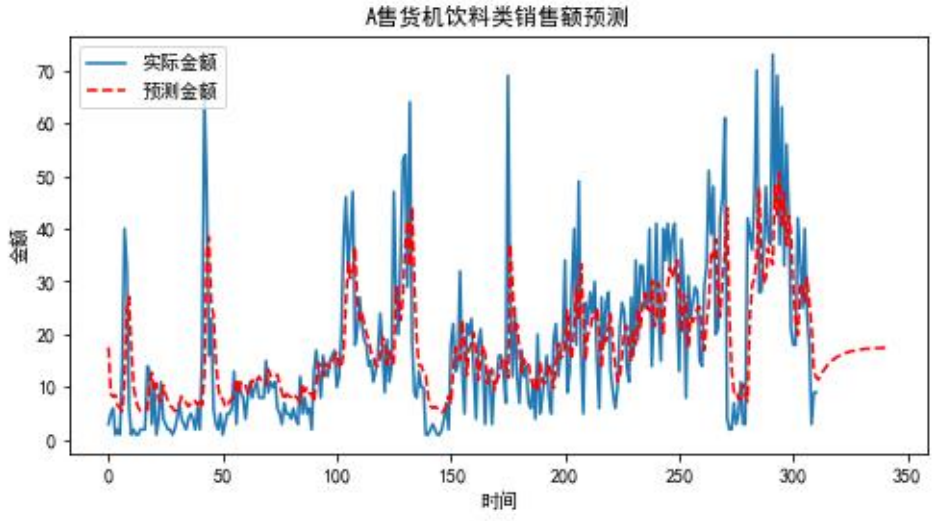


图 4-2-1 A 售货机饮料类商品销售额预测

由图 4-2-1 可以看出，ARMA 模型预测得出的拟合效果较好，预测金额与实际金额基本相似，但一些峰值数据拟合效果不佳。

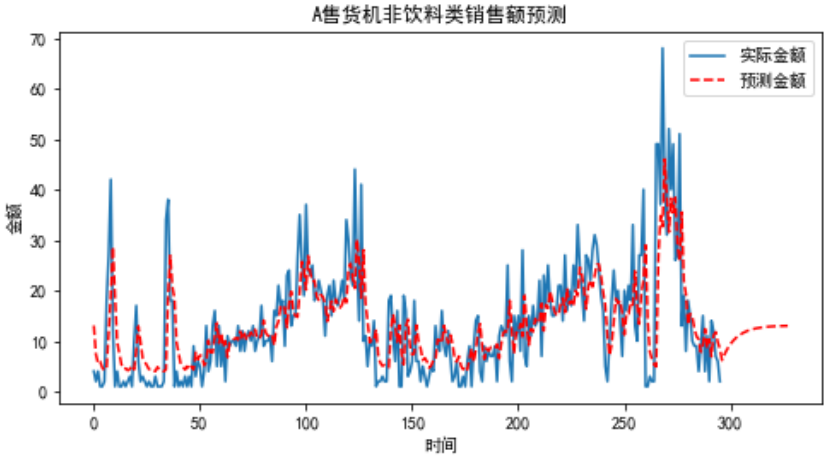


图 4-2-2 A 售货机非饮料类商品销售额预测

预测效果与饮料类相似。该模型预测得出的 A 售货机饮料类商品 1 月份销售额为 505.91 元，非饮料类商品 1 月份销售额为 364.67 元。

其它地区售货机处理过程与 A 售货机相同，所得结果如下。

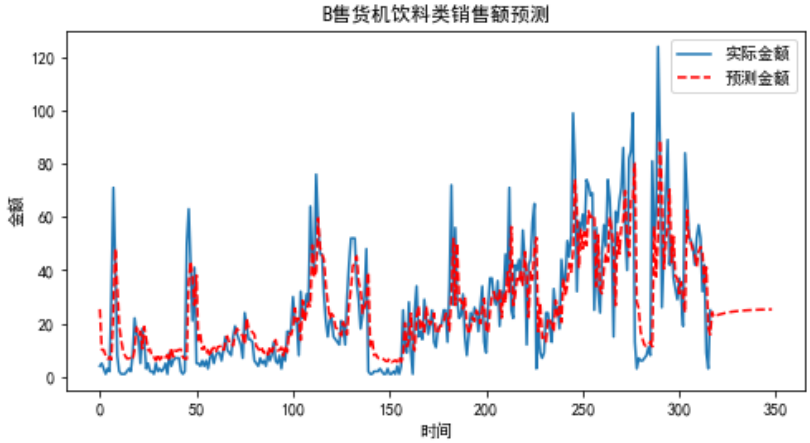


图 4-2-3 B 售货机饮料类商品销售额预测

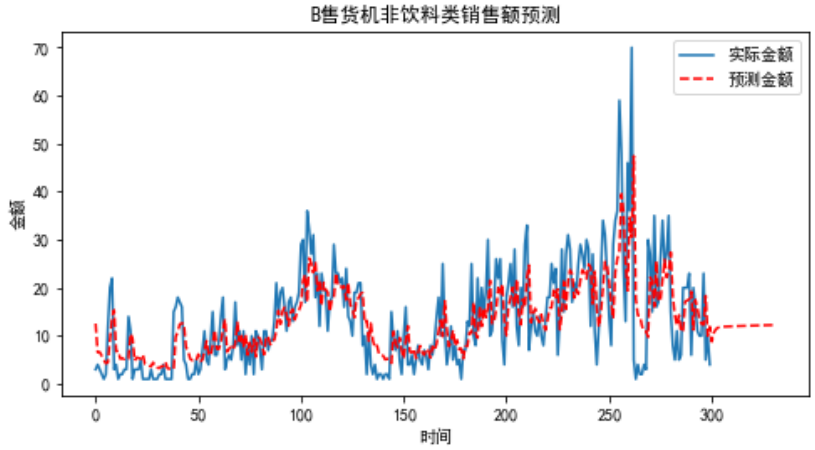


图 4-2-4 B 售货机非饮料类商品销售额预测

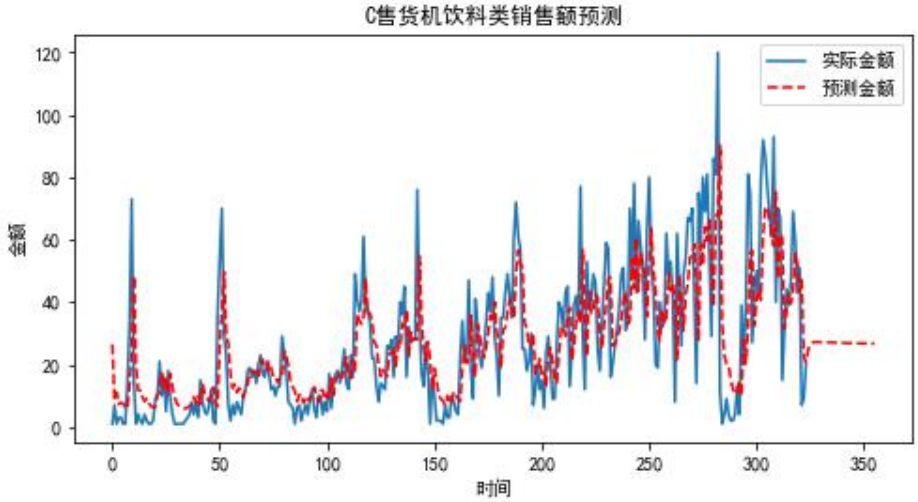


图 4-2-5 C 售货机饮料类商品销售额预测

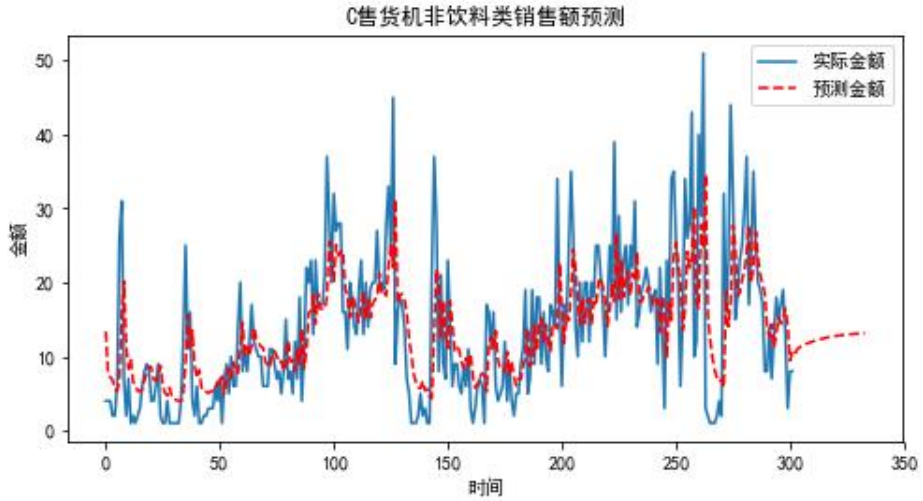


图 4-2-6 C 售货机非饮料类商品销售额预测

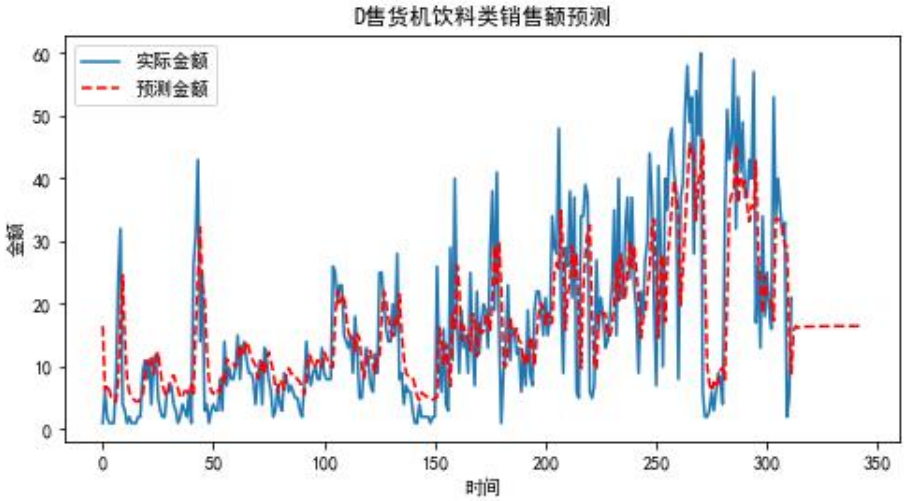


图 4-2-7 D 售货机饮料类商品销售额预测

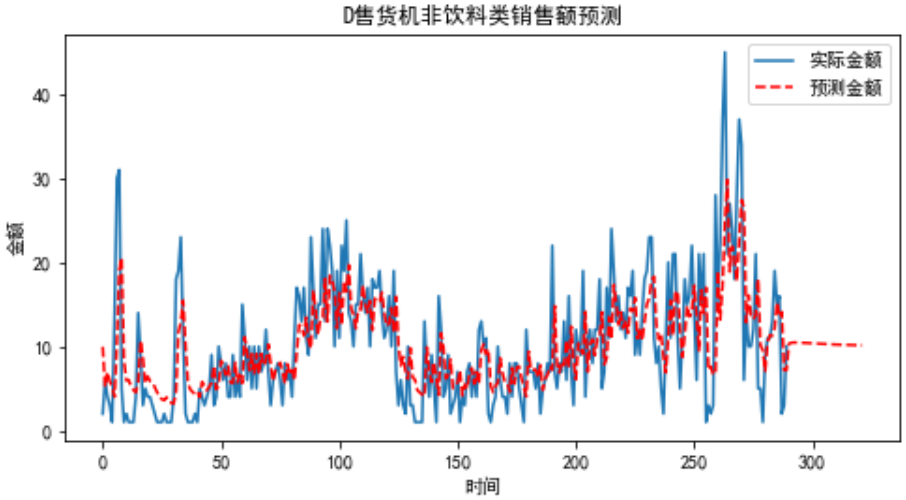


图 4-2-8 D 售货机非饮料类商品销售额预测

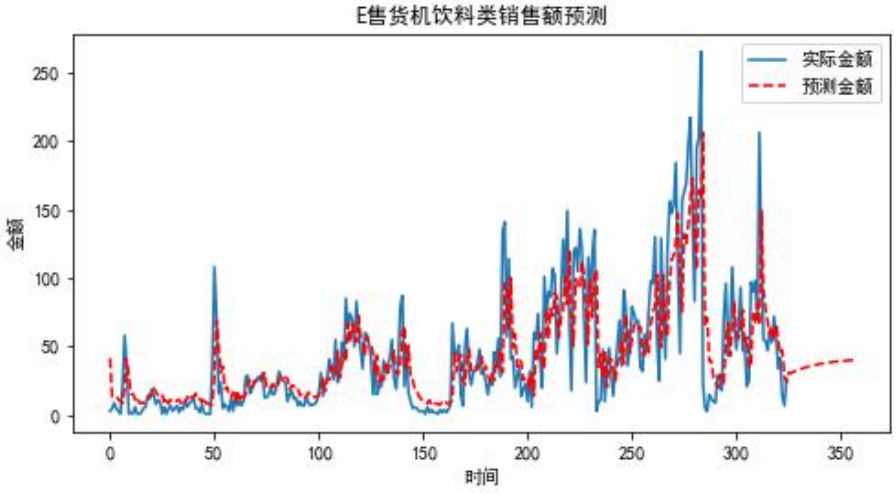


图 4-2-9 E 售货机饮料类商品销售额预测

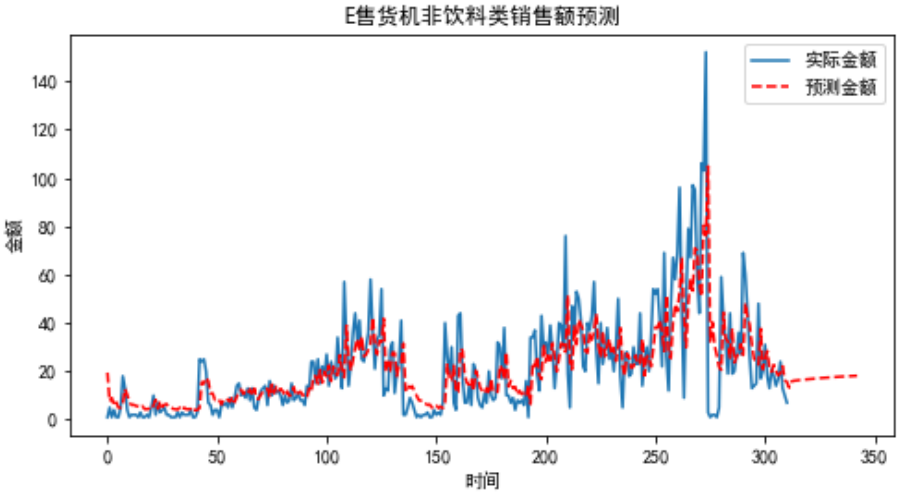


图 4-2-10 E 售货机非饮料类商品销售额预测

4.3 预测交易额

经过上述过程的模型训练与预测，所得 1 月份每台售货机的每个大类商品的交易额如表 4-3-1.

表 4-3-1 预测所得交易额（单位：元）

	A	B	C	D	E
饮料类销售额	505.914	761.044	833.53	509.693	1143.407
非饮料类销售额	364.674	371.989	386.571	318.972	533.927