自动售货机项目报告

一、项目背景

自动售货机是当前比较便利的一种新零售体验，为传统经营模式节省人力和财力成本。但自动售货机商品供应种类、频率、供应量等问题是经营者持续关注的问题。本项目通过对自动售货机商品的数据分析能帮助经营者掌握商品销售情况，是及时掌握经营方向的有效手段之一。

本项目提供了 2017 年 1 月 1 日至 2017 年 12 月 31 日的某商场不同地点的 5 台自动售货机销售数据，五台售货机编号分别为 A、B、C、D、E，附件一是商品销售数据，附件二是商品类别。

数据量：附件一 70680 行，9 列；附件二 315 行，3 列。

数据属性如表 1.1 所示。

表 1.1 原始数据属性

属性	描述
订单号	购买商品的订单号，有可能购买多个同商品但订单号只有一个
设备 ID	自动售货机 ID，分别对应 A/B/C/D/E
应付金额	商品标签价格
实际金额	用户实际支付金额
商品	具体的商品名称
支付时间	消费者订单交易时间
地点	A/B/C/D/E
状态	是否出货、是否退款，正常状态：已出货未退款
提现	商品是否提现
大类	商品的一级分类，饮料类/非饮料类
二级类	商品的二级分类

二、数据预处理与分析

在对数据进行分析之前，先要对原始数据进行预处理，包括检查原始数据的有效性和一致性等。附件一中最后一行支付日期 2 月 29 日不存在，故将其删掉。支付时间列的数据类型从 object 类修改为 datetime 类。针对同一商品出现不同

价格，考虑商品促销减付、商家联合支付软件搞的满减/免等因素，均视为正常情况，这里不做处理。原始数据无冗余值、无缺失值。

1. 任务 1.1

内容：根据附件 1 中的数据，提取每台售货机对应的销售数据，保存在 CSV 文件中。

提取每台售货机销售数据在数据预处理之后完成，主要使用 query 方法即可；在保存时，注意保存编码为 utf_8_sig，使用 utf_8_sig 中文会出现乱码。保存文件见 task1_A.SCV、task1_B.SCV、task1_C.SCV、task1_D.SCV、task1_E.SCV。

2. 任务 1.2

内容：计算每台售货机 2017 年 5 月份的交易额、订单量及所有售货机交易总额和订单总量。

获取 5 月份数据可将“支付时间”列设置为行索引，切片取“2017-05”数据即得 5 月份数据，获取交易额即对“实际金额”列使用 sum 方法，获取订单量即对 5 月份数据使用 count 方法。其结果如表 2.1 所示。从表 2.1 可以看出，5 月份售货机交易额：E>C>B>A>D 最多，5 月份订单量：E>B>C>A>D。

表 2.1 每台售货机交易额和订单量

售货机	交易额	订单量
A（5 月份）	3385.1	756
B（5 月份）	3681.2	869
C（五月份）	3729.4	789
D（5 月份）	2392.1	564
E（5 月份）	5699	1292
5 月份所有售货机	18886.8	4270
2017 年所有售货机	286979.7	70679

3. 任务 1.3

内容：计算每台售货机每月的每单平均交易额与日均订单量。

计算每月每单平均交易额即按月分组并用 mean 方法对“实际金额”列求平均；日均订单量是每台售货机订单总量除以 2017 年天数，即 365 天，并对订单量取整数；每月日均订单量即按月分组来统计的订单量除以 31 天， 如表 2.2 所示为每台售货每月每单平均交易额与日均订单量。从表 2.2 看出 E 的日均订单量最多，每台售货机每月每单平均交易额保持在[3, 5]之间，差额较小。

表 2.2 每台售货机每月每单平均交易额与日均订单量

每月每单平均交易额 / 月份	售货机				
	A	B	C	D	E
1	4.506567	3.753005	4.328496	3.692664	4.680226
2	3.864035	3.255676	3.826087	3.088652	3.638372
3	3.58549	3.614717	3.769962	4.305729	4.305714
4	4.036913	4.07529	4.403678	3.790293	4.159888
5	4.477646	4.236133	4.726743	4.241312	4.410991
6	4.047394	4.06805	4.5017	4.025962	3.817856
7	4.097689	4.401739	3.988351	4.229653	3.919311
8	3.358709	3.5842	3.913582	3.316503	3.804471
9	4.307212	4.130258	4.427294	3.89939	4.125375
10	4.020703	4.11234	4.27333	3.884233	3.676125
11	4.471552	4.268784	4.352393	3.862314	4.283227
12	3.787868	3.667014	3.943043	3.57258	4.168973
每月日均订单量	A	B	C	D	E
1	12	12	12	8	11

2	6	6	7	5	8
3	9	9	8	6	11
4	19	19	24	14	29
5	28	28	25	28	42
6	60	60	61	34	84
7	11	11	25	10	26
8	32	32	41	23	57
9	56	56	54	32	133
10	65	65	71	38	90
11	66	66	63	39	162
12	71	71	77	54	105
日均订单量	29	37	40	24	64

三、数据分析与可视化

3.1 任务 2.1

内容：绘制 2017 年 6 月销量前 5 的商品销量柱状图。

柱状图直观反映了商品销量的分布情况，该部分主要是提取 2017 年月份数据，对商品分组统计销量，然后，对数据降序排序取前五行。其得到结果如表 3.1 所示。

表 3.1 2017 年 6 月销量前 5 的商品

商品	2017 年 6 月销售量
怡宝纯净水	657
40g 双汇玉米热狗肠	240
东鹏特饮	238
脉动	235
250ml 维他柠檬茶	225

2017 年 6 月销量前 5 的商品销量柱状图如图 3.1 所示。从图 3.1 可以看出，

销售量第一的是怡宝纯净水，且有断层，40g 双汇玉米热狗肠占销量第二，东鹏特饮占销量第三，脉动占销量第四，250ml 维他柠檬茶占第五。

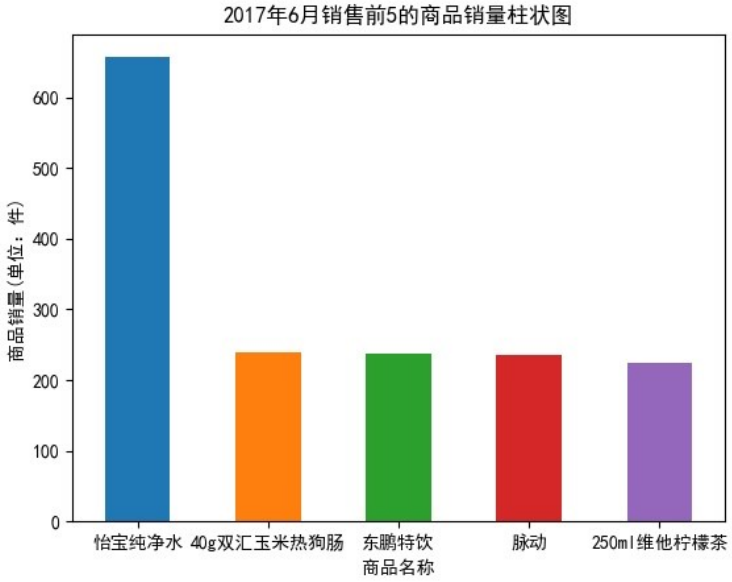

图 3.1 2017 年 6 月销售前 5 的商品销量柱状图

3.2 任务 2.2

内容：绘制每台售货机每月总交易额折线图及交易额环比增长率柱状图。

折线图反应了每台售货机每月交易额的变化趋势，对每台售货机按月分组计算“实际金额”的和即得每月总交易额，得到结果如表 3.2 所示。

表 3.2 每台售货机每月交易额

每月交易额 / 月份	售货机				
	A	B	C	D	E
1	1509.7	1373.6	1640.5	956.4	1656.8
2	440.5	602.3	792.0	435.5	938.7
3	914.3	957.9	991.5	826.7	1507.0
4	1804.5	2457.4	3232.3	1679.1	3723.1
5	3385.1	3681.2	3729.4	2392.1	5699.0
6	6755.1	7550.3	8472.2	4187.0	9899.7
7	1950.5	1518.6	3047.1	1340.8	3186.4

8	2236.9	3516.1	4927.2	2371.3	6722.5
9	4479.5	7207.3	7429.0	3833.1	17054.3
10	6292.4	8331.6	9469.7	4606.7	10208.6
11	5187.0	8669.9	8456.7	4673.4	21501.8
12	7587.1	8104.1	9380.5	5941.2	13557.5

每台售货机每月总交易额折线图如图 3.2 所示。从图可看出，E 售货机总体每月交易额领先其它售货机，且在 6、9、11 月达到极值，11 月份达到最大值，C 售货机次之，D 的每月交易额最差。

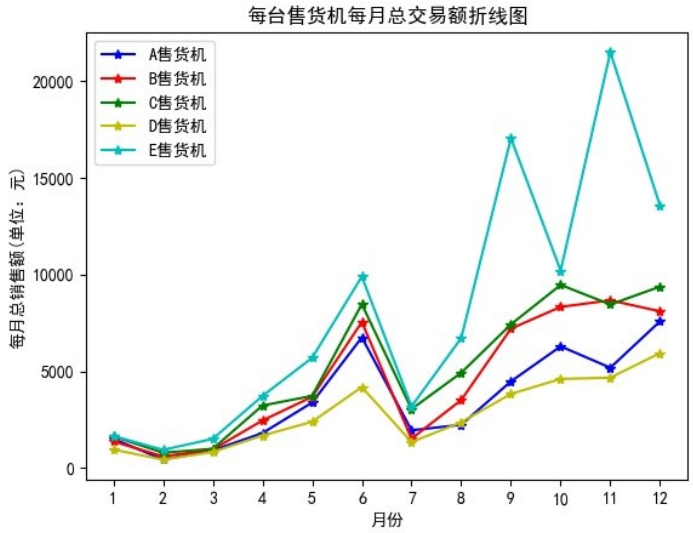

图 3.2 每台售货机每月总交易额折线图

交易额环比增长率=(本期交易额-上期交易额)/上期交易额，反应的是连续两月的交易额变化。由表 3.2 得到的数据通过 pct_change 方法可得到每月交易额环比增长率，每台售货机每月交易额环比增长率柱状图如图 3.3 所示。因为第一个月是首月，其上一期交易额为 NaN，所以没有增长率，从图 3.3 可以看出，C 售货机 3 月的增长率是最高的；在 2、7 月份，所有售货机环比增长率为负，交易额相比上一期有明显下降。

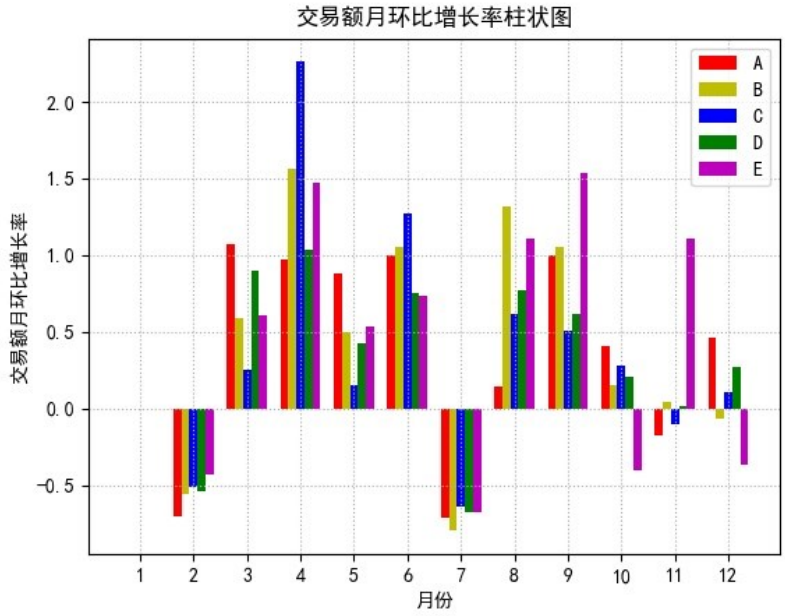

图 3.3 每台售货机每月交易额环比增长率柱状图

3.3 任务 2.3

内容：绘制每台售货机毛利润占总毛利润比例的饼图（假设饮料类毛利率为 25%，非饮料类为 20%）。

饼图反映毛利率比例的百分比情况，每台售货机毛利润占总毛利润比例=(每台售货机饮料类实际金额之和×25%+每台售货机非饮料类实际金额之和×20%)/(所有售货机饮料类实际金额之和×25%+所有售货机非饮料类实际金额之和×20%)，如图 3.4 是每台售货机毛利润占总毛利润比例的饼图。从图 3.4 可看出，总毛利润主要来自 E 毛利润，C 次之，D 毛利润最少。

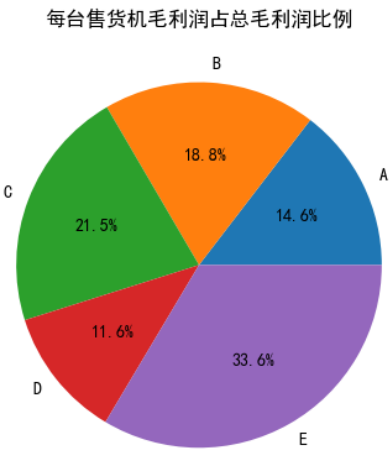

图 3.4 每台售货机毛利润占总毛利润比例的饼图

3.4 任务 2.4

内容：绘制每月交易额均值气泡图，横轴为时间，纵轴为商品的二级类目。

气泡图反映时间、商品二级类目和每月交易额均值三个变量之间的关系，其中将每月交易额均值作为气泡图大小，时间取的是月份，气泡图可以直观呈现商品的每月交易额相对大小。每月交易额均值数据见附件“每月交易额均值.csv”。

每月交易额均值即对“二级类”、“月份”分组再计算“实际金额”的均值。如图 3.5 所示为每月交易额均值气泡图，由于每月平均交易额均值较小，气泡不明显，所以将均值扩大了 15 倍，以便于观察气泡图。从图 3.5 可以看出，每个月各个商品交易额均值变化趋势、每个商品各个月交易额相对大小；其中，每个月的“香烟”气泡都最大，表明它的每月交易额相比其它二级类目多，图中没有气泡的点表示当月该商品交易额为 0。

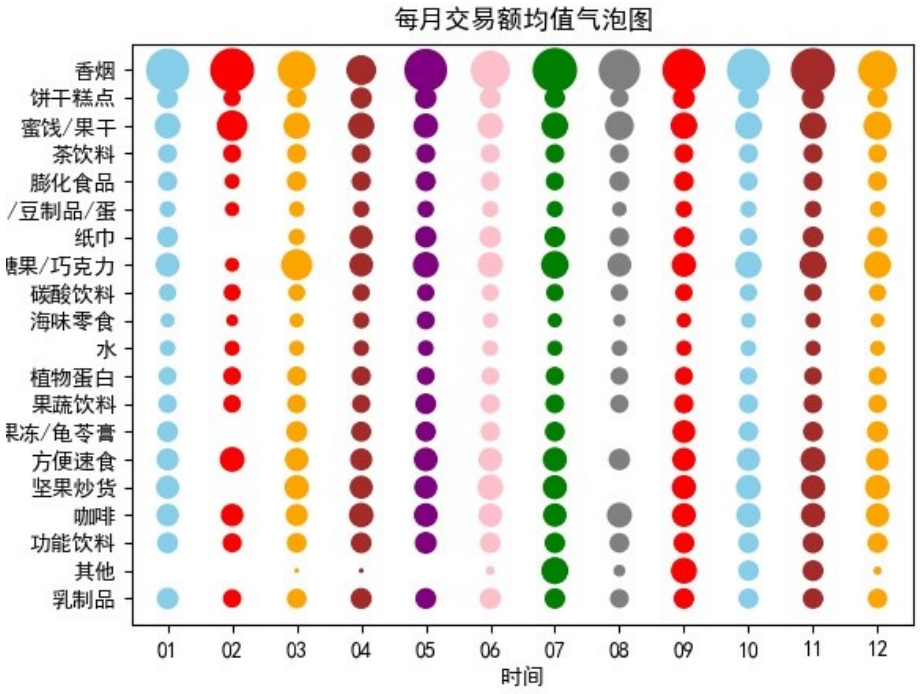

图 3.5 每月交易额均值气泡图

3.5 任务 2.5

内容：绘制售货机 C 6、7、8 三个月订单量的热力图，横轴以天为单位，纵轴以小时为单位，从热力图可以分析得出哪些结论？

热力图反映在哪天哪个小时的订单密集程度，也就是订单量相对多或少。6 月份订单量的热力图如图 3.6 所示，7 月份订单量的热力图见图 3.7，8 月份订单量的热力图见图 3.8。从图 3.6 可以看出，6 月份订单量前三的依次是：23 日 16 时、25 日 16 时，27 日 16 时，都集中在 16 时。图 3.7 看出 7 月份订单量前三和 6 月份一样；图 3.8 表明，8 月份订单量第一的是 18 日 16 点，19 日 16 点次之，18 日 17 点第三。总的来说，6、7、8 月份订单量均在 16 点是最多的，订单量较多的均集中在后半个月，订单量普遍集中在早上 8 点到晚上 22 点，6 月份订单量较密集且较多，7 月份订单量较稀疏且较小。

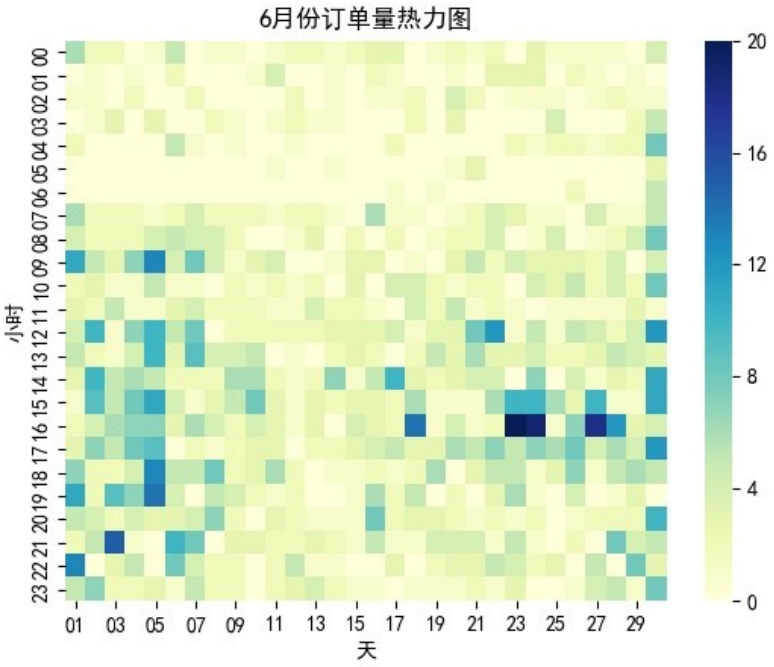

图 3.6 6 月份订单量热力图

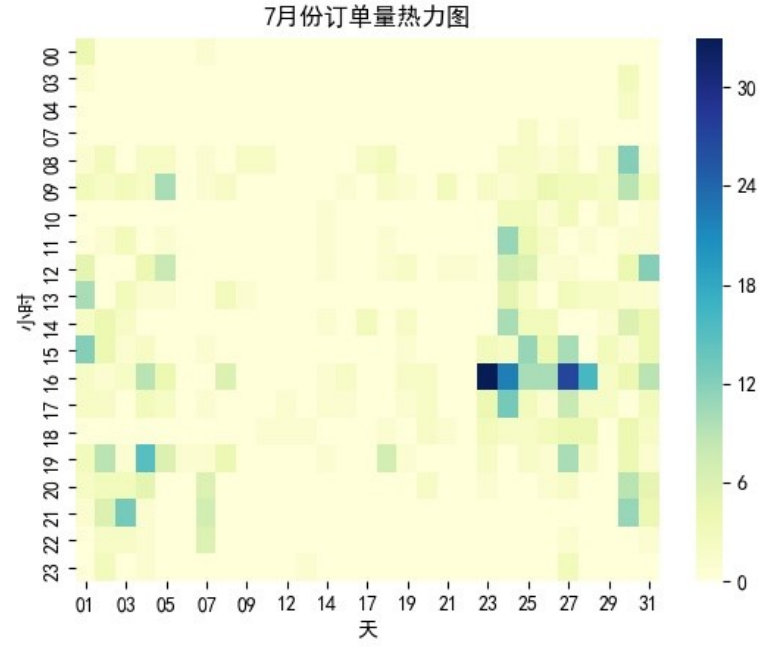

图 3.7 7 月份订单量热力图

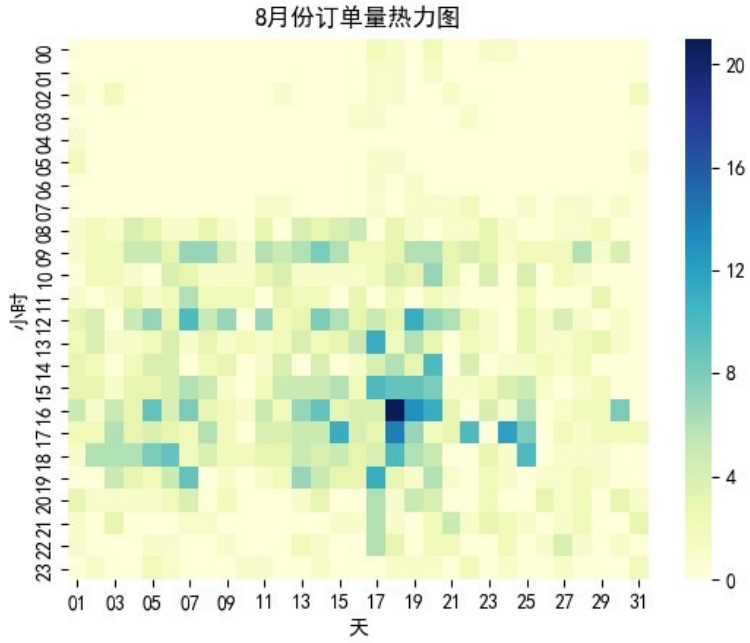

图 3.8 8 月份订单量热力图

四、自动售货机肖像

4.1 任务 3.1

内容分析各售货机商品销售数据，总结规律，给出每台售货机饮料类商品的标签，并保存在文件中。

标签能直观反映商品某一特征的特点。从表 4.1 可以看出，每台售货机 25%、50%、75%的销量数值都较小，不适合作为标签阈值。根据数据分布情况分析，商品标签划分结合表 4.1 及考虑销量和销售连续性，定义滞销商品为小于销量后 25%的且没有月转的商品（即只销售一个月），热销商品为销量前 5%且有连续月转的商品（即连续销售 12 个月），其余为正常销售。

表 4.1 每个售货机数据描述

售货机	describe()
	count 112.000000 mean 54.678571 std 86.686264 min 1.000000

A	25%	6.000000
	50%	19.000000
	75%	67.250000
	max	531.000000
B	count	115.000000
	mean	78.539130
	std	157.267472
	min	1.000000
	25%	5.000000
	50%	17.000000
	75%	94.500000
	max	1342.000000
C	count	114.000000
	mean	86.043860
	std	152.861288
	min	1.000000
	25%	6.250000
	50%	24.000000
	75%	79.750000
	max	999.000000
D	count	105.000000
	mean	52.495238
	std	87.096777
	min	1.000000
	25%	3.000000
	50%	15.000000
	75%	64.000000
	max	411.000000
E	count	113.000000
	mean	141.283186
	std	237.236187
	min	1.000000
	25%	11.000000
	50%	47.000000
	75%	169.000000
	max	1705.000000

其结果见附件“task3_1A.csv”、“task3_1B.csv”、“task3_1C.csv”、“task3_1D.csv”、“task3_1E.csv”，数据属性见表 4.2。

表 4.2 生成商品标签后的数据属性

属性	描述

饮料类商品	每个饮料类商品名称
标签	滞销/正常/热销

4.2 任务 3.2

内容：扩展标签，依据标签生成完整的售货机画像。

考虑根据应付金额增加消费档次标签，由销售价格从高到低将消费档次分为高档/中档/低档。低档定义为价格<销售价格的 25%，高档定义为价格>销售价格的 75%，其它为中档，理论上按商品单价来划分消费档次，但这里并不知道单价，而且由于促销原因(促销力度不知道)降价不能算出每单销售数量，所以不考虑。划分时可能存在一件商品存在多个价格，从而有多个档次，这里只保留一个，也就是第一次出现的重复值，扩展的标签结果见附件任务 3.2。

画像以可视化图片形式间接明了的直观看出售货机销售特点。该部分对标签进行绘制词云图，词云图将出现频次较高的标签以视觉上突出呈现，对各个售货机销量较好的商品一目了然。词频=每台售货机饮料类每个商品销量/每台售货机总销量，生成的词频数据表明，词频小于 0.01 的数据销量不突出，为了更加清晰的展示售货机特色，可将这部分数据剔除。

各个售货机词云图见图 4-1~4-5 所示。由 4-1~4-5 可知，所有售货机的“怡宝纯净水”销量持高，可以各个地方较大力度的增加其供应量；每个售货机可根据词云图展示的视觉突出性，适当增加其它商品库存，同时减少滞销商品的供应量，具体参考如下：

A 售货机适当增加“东鹏特饮”、“脉动”、“阿萨姆奶茶”、“营养快线”、“统一冰红茶”、“205ml 维他原味豆奶”的库存，减少“健能酸奶”、“天地一号”、“安慕希酸奶”、“小茗同学冷泡茶（乳酸菌味）”、拿铁咖啡（统一）、“脉动（椰子菠萝口味）的供应量；

B 售货机适当增加“东鹏特饮”、“脉动”、“阿萨姆奶茶”、“营养快线”、“统一冰红茶”、“王老吉(500ml)” 的库存”，“减少 250ml 维他椰子植物蛋白饮料 9374 ”、“500ml 恒大冰泉矿泉水”、“商品 1”、“商品 14”、“商品 2”、“娃哈哈红枣酸奶”、“小茗同学”、“拿铁咖啡（统一）”、“脉动（椰子菠萝口味）”、“雪碧”

的供应量；

C 适当增加“脉动”、“东鹏特饮”、“营养快线”、“阿萨姆奶茶”、“王老吉(罐)”、“统一冰红茶”的库存，减少“250ml 维他椰子植物蛋白饮料 9374”、“商品 14”、“拿铁咖啡（统一）”、“维他奶黑豆奶饮品”、“芦荟汁” 的供应量；

D 适当增加 “营养快线”、“阿萨姆奶茶”、“东鹏特饮”、 “脉动” 、“统一冰红茶”、“统一绿茶”、“果粒橙” 的库存，减少 “145ml 旺仔牛奶罐装”、“250ml 维他椰子植物蛋白饮料 9374”、“250ml 香满楼纯牛奶”、“500ML 宝矿力”、“500ml 恒大冰泉矿泉水”、“商品 1”、“商品 2”、“小茗同学冷泡茶（乳酸菌味）”、“旺仔牛奶 145ml”、“经典奶茶”、“芦荟汁”、“脉动（青柠）”、“脉动（椰子菠萝口味）” 的供应量；

E 适当增加 “营养快线”、“脉动”、“东鹏特饮”、“阿萨姆奶茶”、“统一冰红茶”、“统一绿茶”、“果粒橙”、“王老吉”，减少 “商品 2”、“小茗同学冷泡茶（乳酸菌味）”、“小茗同学（青柠红茶）”、“珠江纯生啤酒”、“维他奶黑豆奶饮品” 的供应量。

图 4-1 A 自动售货机词云图

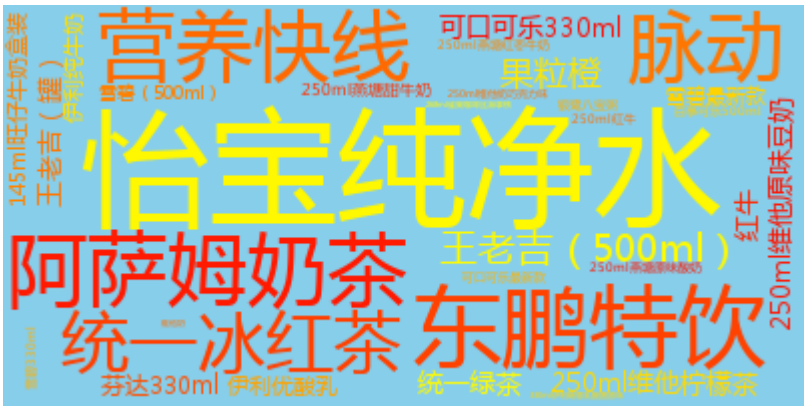

图 4-2 B 自动售货机饮料类商品词云图

图 4-3 C 自动售货机饮料类商品词云图

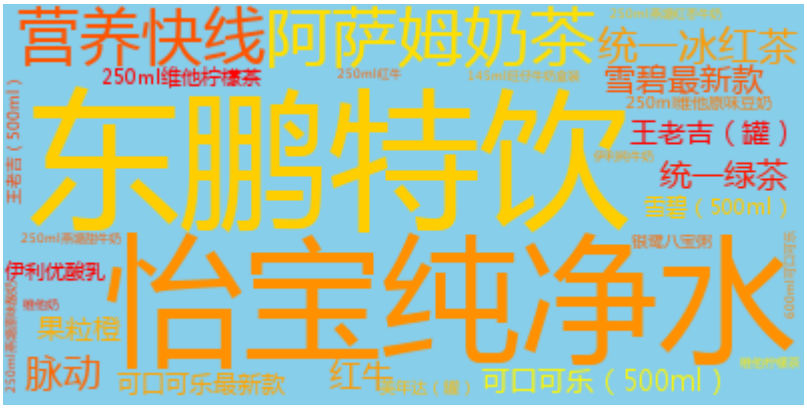

图 4-4 D 自动售货机饮料类商品词云图

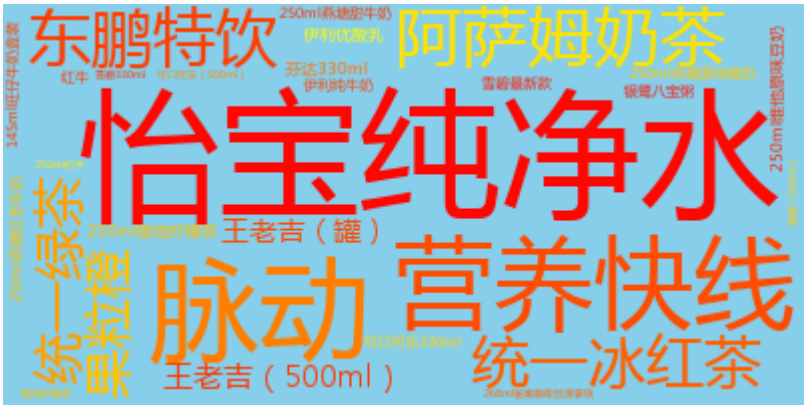

图 4-5 E 自动售货机饮料类商品词云图

五、业务预测

预测未来销售额的原理：将历史销售数据的销售额（即实际金额）作为目标值，其他列属性作为特征值，并合理划分数据为训练集和测试集，并将训练集放入机器学习模型中学习，然后将测试集用作验证学习效果，最后调参优化结果。

不能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。由附件数据可以衍生出月销量属性，图 5.1 展示的是 B 售货机月销量月与实际金额(即每月交易额)的关系，可以看出每月交易额与月销量呈正比，其它售货机同上，理论上可以作为预测的特征之一，但是得到的数据很少，一年内每个大类月交易额只有 12 个数据，不足以支撑预测数据量，而且附件提供的列属性，只有少部分有使用价值，需要经营者提供更多数据，同时提供商品更多维度，比如：商品单价(可以根据单价来判断消费档次)，促销情况(促销力度会影响客户购买商品)、客户满意度(商品的综合评价会影响客户购买)、商品热量情况(高脂/低脂/正常)、商场周边环境(比如居民楼偏多或办公场所偏多各自的需求不一样)等维度。

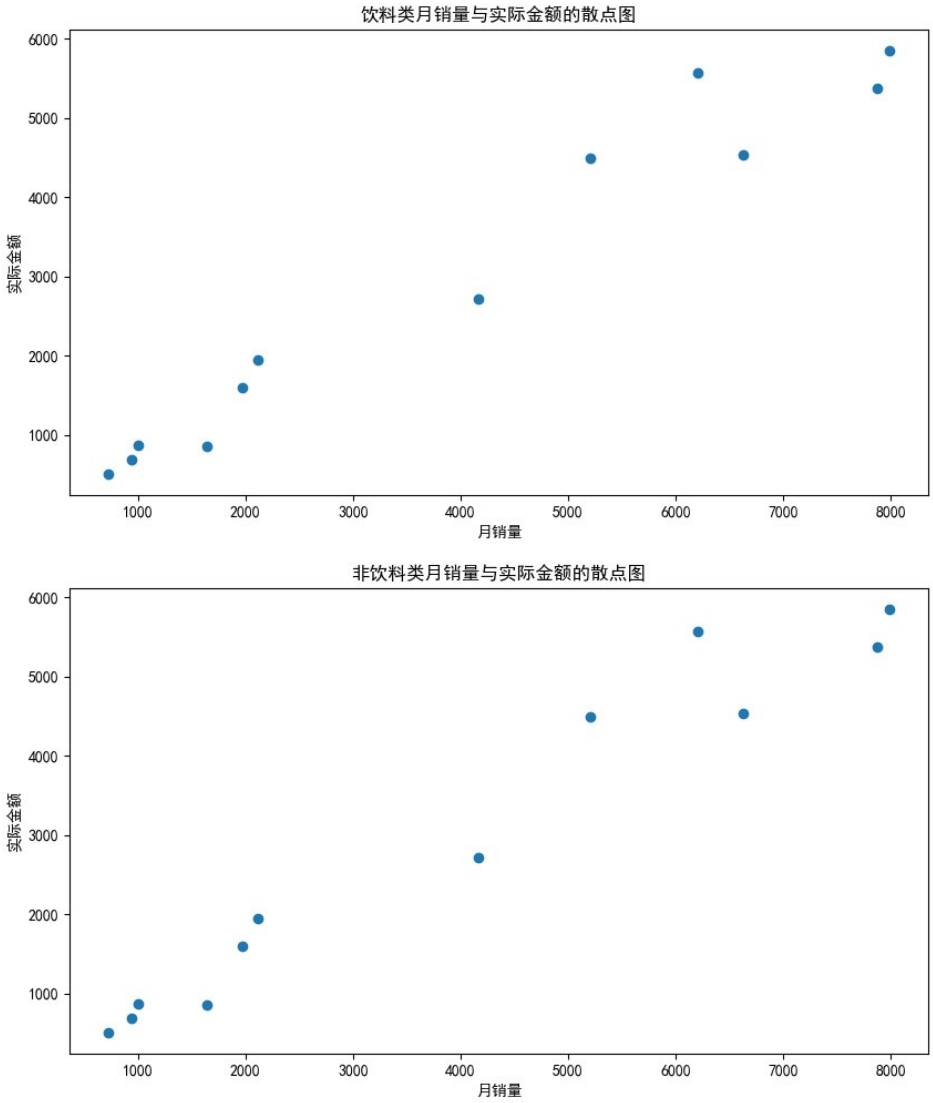

图 5.1 月销量与实际金额的散点图