本篇报告旨在分析 ABCD4 台自动售货机在 2017 年的商品销售情况

# 一、 数据预处理

先做数据清洗：

**数据清理**

数据清理(data cleaning) 的主要思想是通过填补缺失值、光滑噪声数据，平滑或删除离群点，并解决数据的不一致性来“清理“数据。如果用户认为数据时脏乱的，他们不太会相信基于这些数据的挖掘结果，即输出的结果是不可靠的。

处理缺失值：删除变量，填充变量

处理异常值：将一些明显异常（不合理）的值进行删除或更改

在本次数据分析中我们做以下处理：

1. 处理异常事件序列数据（删除）：检测出有一条记录的时间数据异常，并删去
2. 检查同种商品是否有订单价格不一致的异常情况：无
3. 检测“应付金额”与“实际金额”是否有区别：无

| 每台售货机 2017 年 5 月的交易额、订单量 | | | |
|---|---|---|---|
| | 订单总量 | 交易总额 | 每单均值 |
| A_May | 756 | 3385.1 | 4.48 |
| B_May | 869 | 3681.2 | 4.24 |
| C_May | 789 | 3729.4 | 4.73 |
| D_May | 564 | 2392.1 | 4.24 |
| E_May | 1292 | 5699 | 4.41 |
| 总计 | 4270 | 18886.8 | 4.42 |

首先让我们来看看所有售货机的订单量交易总额与均值：

| 所有售货机交易总额和订单总量 | | | |
|---|---|---|---|
| | 订单总量 | 交易总额 | 每单均值 |
| A | 10486 | 42542.6 | 4.06 |
| B | 13482 | 53970.3 | 4 |
| C | 14493 | 61568.1 | 4.25 |
| D | 8713 | 33243.3 | 3.82 |
| E | 23505 | 95655.4 | 4.07 |
| 总计 | 70679 | 286979.7 | 4.06 |
| | | | |

直接从数值分析我们可以看到订单最多的是售货机 E，但每单均值最高的是 C，说明 C 的消费者消费能力较强，倾向买价格更高的商品；然后订单数最少的是 D，每单均值最少的也是 D 只有 3.82 元/单。

计算每台售货机每月的每单平均交易额与日均订单量，做成表格：

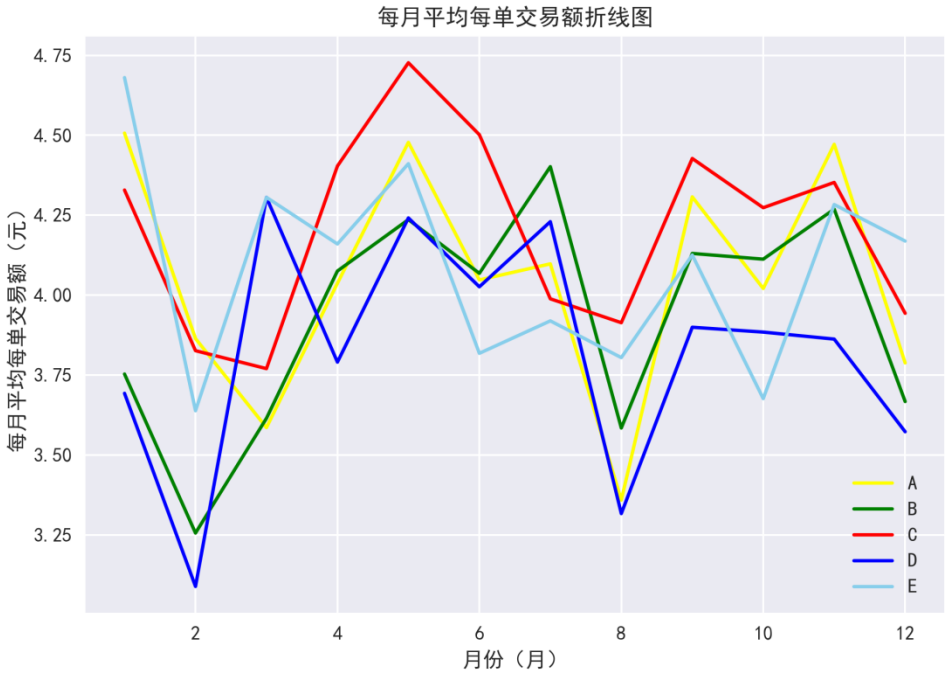
| | A | | B | | C | | D | | E | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 订单总额 | 日均订单量 | 订单总额 | 日均订单量 | 订单总额 | 日均订单量 | 订单总额 | 日均订单量 | 订单总额 | 日均订单量 |
| 1 | 1509.7 | 10.81 | 1373.6 | 11.81 | 1640.5 | 12.23 | 956.4 | 8.35 | 1656.8 | 11.42 |
| 2 | 440.5 | 4.07 | 602.3 | 6.61 | 792.0 | 7.39 | 435.5 | 5.04 | 938.7 | 9.21 |
| 3 | 914.3 | 8.23 | 957.9 | 8.55 | 991.5 | 8.48 | 826.7 | 6.19 | 1507.0 | 11.29 |
| 4 | 1804.5 | 14.9 | 2457.4 | 20.1 | 3232.3 | 24.47 | 1679.1 | 14.77 | 3723.1 | 29.83 |
| 5 | 3385.1 | 24.39 | 3681.2 | 28.03 | 3729.4 | 25.45 | 2392.1 | 18.19 | 5699.0 | 41.68 |
| 6 | 6755.1 | 55.63 | 7550.3 | 61.87 | 8472.2 | 62.73 | 4187.0 | 34.67 | 9899.7 | 86.43 |
| 7 | 1950.5 | 15.35 | 1518.6 | 11.13 | 3047.1 | 24.65 | 1340.8 | 10.23 | 3186.4 | 26.23 |
| 8 | 2236.9 | 21.48 | 3516.1 | 31.65 | 4927.2 | 40.61 | 2371.3 | 23.06 | 6722.5 | 57.0 |
| 9 | 4479.5 | 34.67 | 7207.3 | 58.17 | 7429.0 | 55.93 | 3833.1 | 32.77 | 17054.3 | 137.8 |
| 10 | 6292.4 | 50.48 | 8331.6 | 65.35 | 9469.7 | 71.48 | 4606.7 | 38.26 | 10208.6 | 89.58 |
| 11 | 5187.0 | 38.67 | 8669.9 | 67.7 | 8456.7 | 64.77 | 4673.4 | 40.33 | 21501.8 | 167.33 |
| 12 | 7587.1 | 64.61 | 8104.1 | 71.29 | 9380.5 | 76.74 | 5941.2 | 53.65 | 13557.5 | 104.90 |

| | 平均每单交易额 | | | | |
|---|---|---|---|---|---|
| 月份 | A | B | C | D | E |
| 1 | 4.51 | 3.75 | 4.33 | 3.69 | 4.68 |
| 2 | 3.86 | 3.26 | 3.83 | 3.09 | 3.64 |
| 3 | 3.59 | 3.61 | 3.77 | 4.31 | 4.31 |

| | | | | | |
|---|---|---|---|---|---|
| 4 | 4.04 | 4.08 | 4.40 | 3.79 | 4.16 |
| 5 | 4.48 | 4.24 | 4.73 | 4.24 | 4.41 |
| 6 | 4.05 | 4.07 | 4.50 | 4.03 | 3.82 |
| 7 | 4.10 | 4.40 | 3.99 | 4.23 | 3.92 |
| 8 | 3.36 | 3.58 | 3.91 | 3.32 | 3.80 |
| 9 | 4.31 | 4.13 | 4.43 | 3.90 | 4.13 |
| 10 | 4.02 | 4.11 | 4.27 | 3.88 | 3.68 |
| 11 | 4.47 | 4.27 | 4.35 | 3.86 | 4.28 |
| 12 | 3.79 | 3.67 | 3.94 | 3.57 | 4.17 |

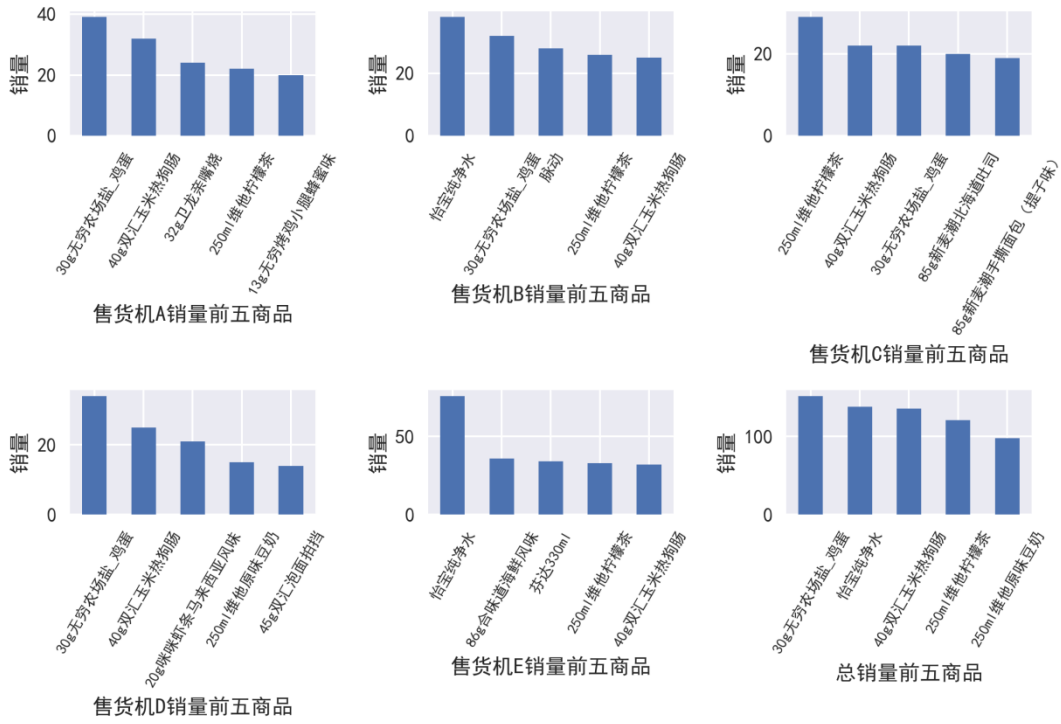
# 二、作图分析

整篇的数据不直观，我们通过折线图观察每月平均交易额的变化以及不同售货机之间的区别



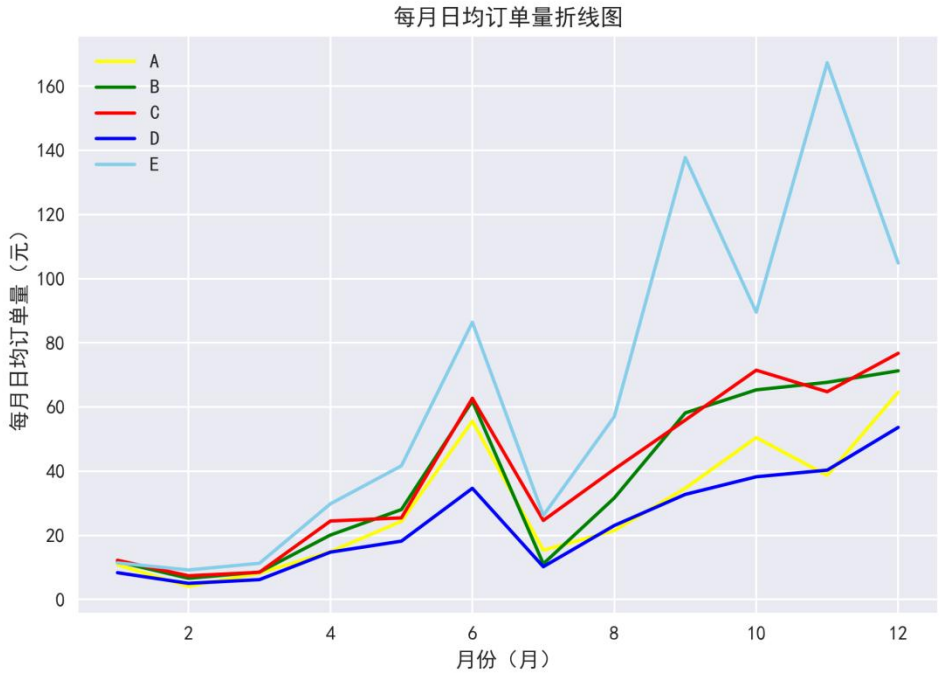

在平均每单交易额上来说 C 基本处于较高位置，D 处于较低位置。从月变化来说 5 台售货机在 5 月份有一个高峰，也就是说在 5 月份价格高的商品卖得更多

那我们来看看 5 月份销量前 5 的商品销量**柱状图**

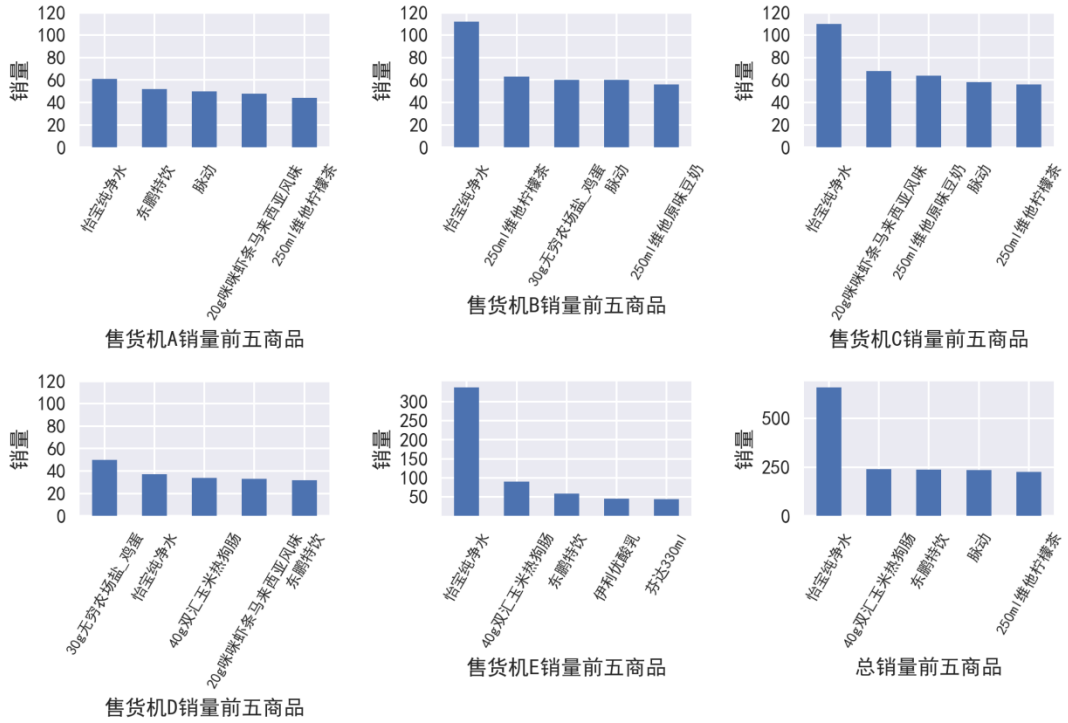



咱们可以看到想一些价格较高的商品比如柠檬茶、鸡蛋、热狗肠、豆奶等商品收人欢迎，但总体的订单量较少。

再来看订单量的变化情况：

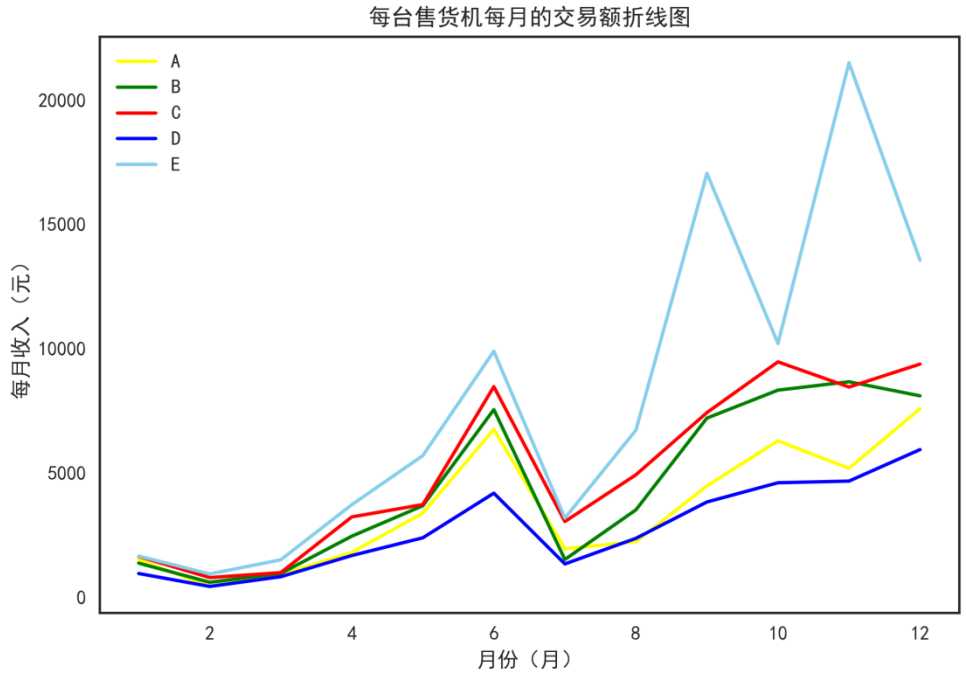


绘制 2017 年 6 月销量前 5 的商品销量**柱状图**

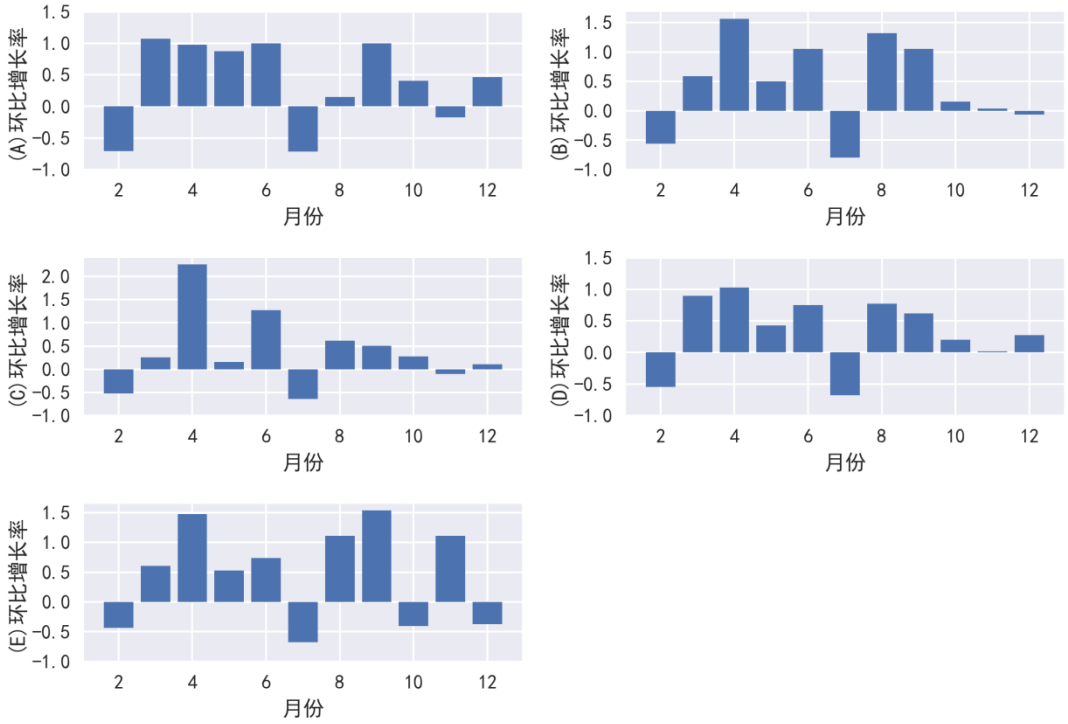



上图可以看到在总体订单量较高的六月的商品销售情况，六月份的怡宝纯净矿泉水非常受欢迎。值得一提的是售货机 E 在此时的日均订单量最大，其中怡宝矿泉水的比重较大。也就是说售货机 E 的消费者对怡宝矿泉水的需求最大，对其他商品不感冒。

绘制每台售货机每月的总交易额**折线图**及交易额月环比增长率**柱状图：**

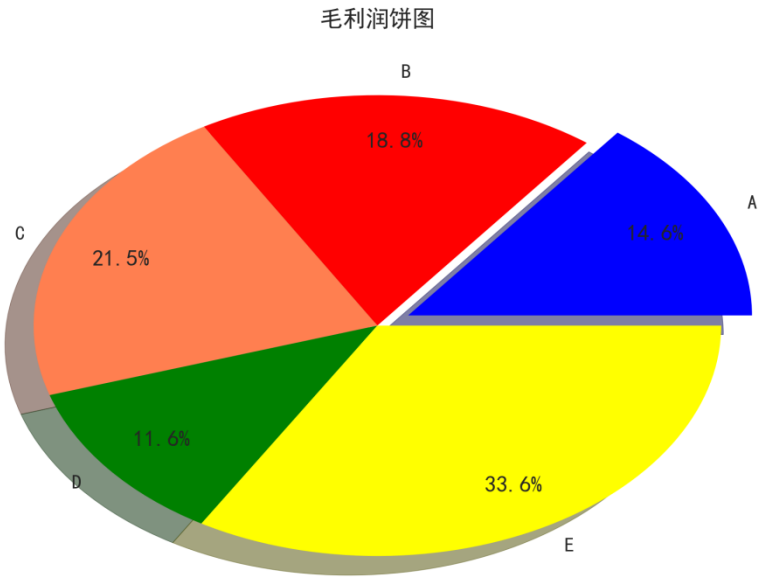


交易额月环比增长率**柱状图：**



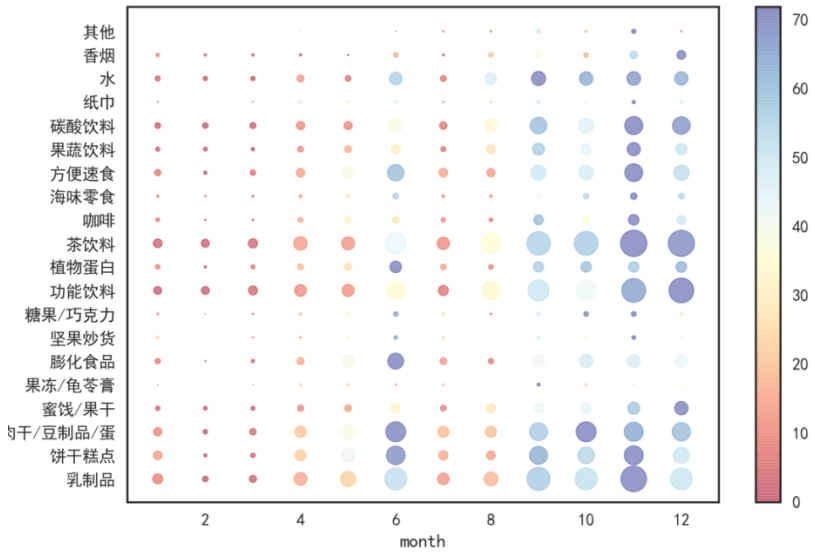

通过上图可以发现 E 的柱状图中每月环比增长率普遍较高，所以交易额增长快。但观察各个图发现在 2 月份和 7 月份的商品交易额普遍下跌，说明此时消费者对售货机商品需求减少。

绘制每台售货机毛利润占毛利润比例的**饼图**



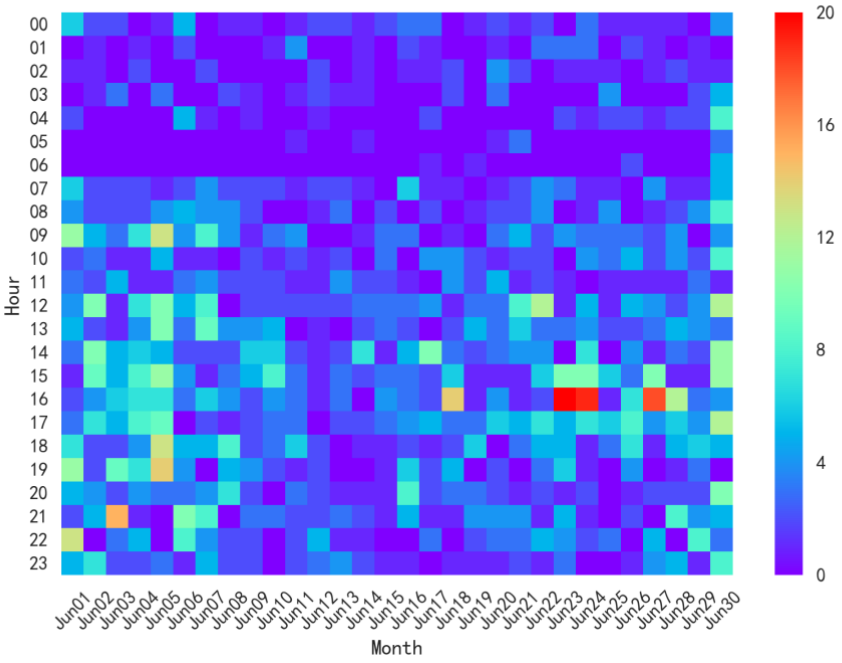

我们可以看到 E 的比重是最高的，“卖得多赚得多”;D 的比重最小，因为卖得少

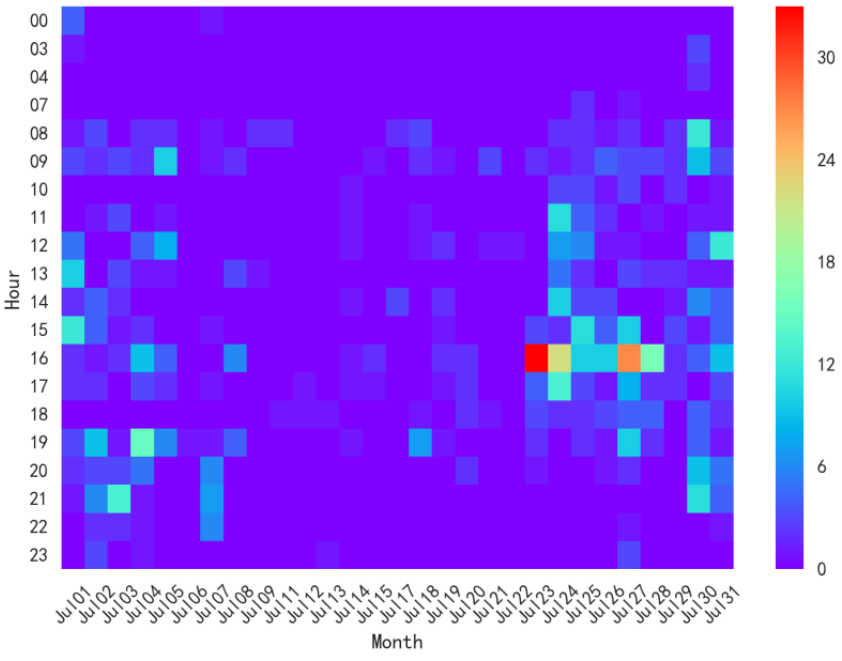
接下来绘制每月交易额均值**气泡图**，横轴为时间，纵轴为商品的二级类目，气泡大小为每月的日交易额均值：

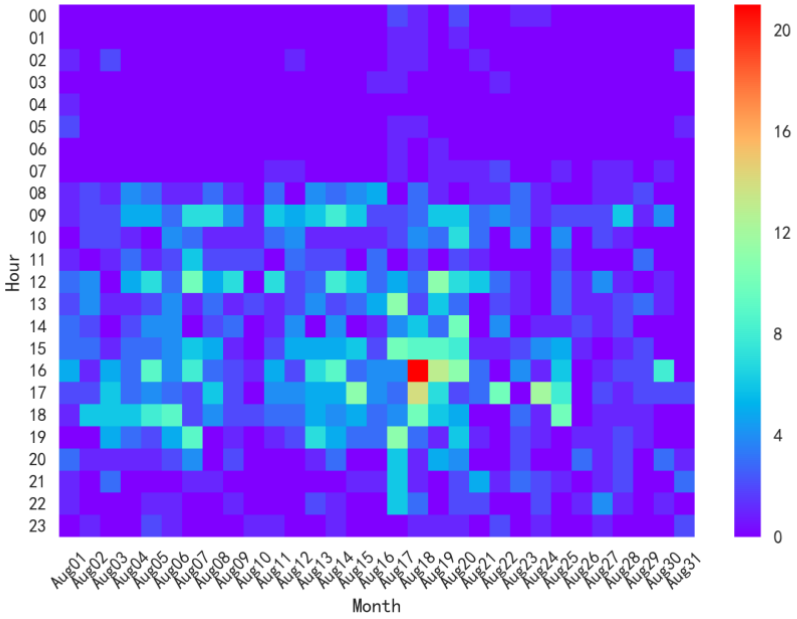



从上图可以看出，在 2017 年 6 月、9、10、11、12 月销量普遍较好，茶饮料与乳制品的销售情况比较好，果冻/龟苓膏、糖果/巧克力、坚果炒货、纸巾等全年销量暗淡

绘制售货机 C6、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。并对热力图进行分析：


Hour
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Jun01 Jun02 Jun03 Jun04 Jun05 Jun06 Jun07 Jun08 Jun09 Jun10 Jun11 Jun12 Jun13 Jun14 Jun15 Jun16 Jun17 Jun18 Jun19 Jun20 Jun21 Jun22 Jun23 Jun24 Jun25 Jun26 Jun27 Jun28 Jun29 Jun30
Month
20
16
12
8
4
0


Hour
00
03
04
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Jul01 Jul02 Jul03 Jul04 Jul05 Jul06 Jul07 Jul08 Jul09 Jul11 Jul12 Jul13 Jul14 Jul15 Jul17 Jul18 Jul19 Jul20 Jul21 Jul22 Jul23 Jul24 Jul25 Jul26 Jul27 Jul28 Jul29 Jul30 Jul31
Month
30
24
18
12
6
0




分析上面热力图发现，消费者主要选择在一天的下午去自动售货机购买商品，而且最高峰往往出现在一个月的中下旬。

# 三、标签与画像

根据商品销售情况我们可以对每台售货机的商品贴上标签以 A 售货机为例：

| | | | |
|---|---|---|---|
| 0 | 怡宝纯净水 | 531 | 热销 |
| 1 | 东鹏特饮 | 448 | 热销 |
| 15 | 雪碧（500ml） | 123 | 热销 |
| 11 | 伊利优酸乳 | 143 | 适中 |
| 29 | 500ml 统一阿萨姆奶茶 | 65 | 适中 |
| 20 | 268ml 雀巢咖啡丝滑拿铁 | 91 | 适中 |
| 19 | 可口可乐（500ml） | 101 | 适中 |
| 17 | 王老吉（500ml） | 108 | 适中 |
| ... | | | |
| 79 | 维他奶黑豆奶饮品 | 7 | 滞销 |
| 80 | 益力多 100g*5 瓶 | 7 | 滞销 |
| 81 | 500ml 营养快线原味 | 6 | 滞销 |
| 55 | 美年达（罐） | 19 | 滞销 |

我首先做第一步分类,按区间[0,90],[90, 360], [>360]分成滞销、适中、热销三个标签，选择这几个分类标签的原因是我认为“热销”应该为日均销售量>1，“适中”在[0.25,1]之间，小于 0.25 可以看做“滞销”;

然后考虑到有一些饮料为 500ml 的，我们应该采取与普通饮料不同的分类标准，于是采用[0,55],[55, 110], [>110]分成滞销、适中、热销三个标签。

其中在做处理时，发现了“益力多 100*5g”与“100*5g 益力多”在附件二里被分为两类的数据处理错误，像这样的还有好几条，暂时找不到好方法解决。

# 画像：

A：怡宝纯净水 东鹏特饮 雪碧（500ml）

B：怡宝纯净水 东鹏特饮 阿萨姆奶茶 脉动 营养快线 王老吉（500ml）

C：怡宝纯净水 脉动 东鹏特饮 阿萨姆奶茶 营养快线 王老吉（罐） 统一冰红茶

D：东鹏特饮、怡宝纯净水 阿萨姆奶茶 可口可乐（500ml） 雪碧(500ml)

E：怡宝矿泉水 脉动 营养快线 阿萨姆奶茶 东鹏特饮 王老吉（500ml

# 四、预测

如果对每台售货机 2018 年的商品销售状况进行预测，我认为现有数据较少，预测效果很差。举个例子，用支持向量机方法预测 A 在 2018 年 1 月的商品销售情况，得出以下结果：

2018 年 1 月 A 商品销售预测:

非饮料 : [1279.11368162]

对比 2017 年 1 月：

792.4

2018 年 1 月 A 商品销售预测:

饮料 : [1468.4636814]

对比 2017 年 1 月：

717.3
差距比较大。
原因为欠拟合，即训练数据过少没能很好地预测
建议先收集更多往年数据，方可进一步通过机器学习进行预测