# 任务 1

## 1.1 提取每台售货机的数据

手动操作 Excel 表格，将售货机 A、B、C、D、E 的数据分在五个文件，分别命名为 task1-1A.csv（以此类推…），并发现 C 售货机中有异常数据（日期为 2017 年 2 月 29 日，手动予以剔除）。再对各个售货机的数据进行去重处理，发现没有其他异常。

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 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 | | | | |
| 2 | DD201708 | E43A6E07 | 3 | 3 | 40g双汇玉 | ###### | A | 已出货未 | 已提现 | | | | |
| 3 | DD201708 | E43A6E07 | 5.8 | 5.8 | 100g卫龙 | ###### | A | 已出货未 | 已提现 | | | | |
| 4 | DD201708 | E43A6E07 | 0.8 | 0.8 | 咪咪虾条 | ###### | A | 已出货未 | 已提现 | | | | |
| 5 | DD201708 | E43A6E07 | 3 | 3 | 250ml维他 | ###### | A | 已出货未 | 已提现 | | | | |
| 6 | DD201708 | E43A6E07 | 3.5 | 3.5 | 东鹏特饮 | ###### | A | 已出货未 | 已提现 | | | | |
| 7 | DD201708 | E43A6E07 | 3 | 3 | 250ml维他 | ###### | A | 已出货未 | 已提现 | | | | |
| 8 | DD201708 | E43A6E07 | 3 | 3 | 卫龙大面 | ###### | A | 已出货未 | 已提现 | | | | |
| 9 | DD201708 | E43A6E07 | 3.5 | 3.5 | 东鹏特饮 | ###### | A | 已出货未 | 已提现 | | | | |
| 10 | DD201708 | E43A6E07 | 1.5 | 1.5 | 卫龙亲嘴 | ###### | A | 已出货未 | 已提现 | | | | |
| 11 | DD201705 | E43A6E07 | 4.5 | 4.5 | 阿萨姆奶 | ###### | A | 已出货未 | 已提现 | | | | |
| 12 | DD201708 | E43A6E07 | 4 | 4 | 145ml旺仔 | ###### | A | 已出货未 | 已提现 | | | | |
| 13 | DD201705 | E43A6E07 | 2 | 2 | 伊利优酸 | ###### | A | 已出货未 | 已提现 | | | | |
| 14 | DD201708 | E43A6E07 | 4.5 | 4.5 | 2g韩国海 | ###### | A | 已出货未 | 已提现 | | | | |
| 15 | DD201708 | E43A6E07 | 4 | 4 | 脉动 | ###### | A | 已出货未 | 已提现 | | | | |
| 16 | DD201708 | E43A6E07 | 3 | 3 | 250ml维他 | ###### | A | 已出货未 | 已提现 | | | | |
| 17 | DD201708 | E43A6E07 | 2 | 2 | 75g新麦滇 | ###### | A | 已出货未 | 已提现 | | | | |

## 1.2 计算售货机的交易额、订单量

先读取数据，然后利用 to_datatime 将提取出相应售货机的支付时间一列数据从字符串格式转换成 datatime 格式。

根据需求，再利用逻辑语句条件将支付时间限制在 2017 年 5 月内，交易总额即实际金额列的总和，订单量即订单号一列的行数。

然后再将得到的数据输入到新建立的 data（dataframe 格式）内，修改 data 的索引然后输出到 csv 文件格式。由于计算的月份不同，所以直接将得到的关于所有售货机的订单总量和交易总额数据手动输入到刚刚生成的 csv 内。

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | 5月的交易 | 5月的订单量 | |
| 2 | A | 3385.1 | 756 | |
| 3 | B | 3681.2 | 869 | |
| 4 | C | 3729.4 | 789 | |
| 5 | D | 2392.1 | 564 | |
| 6 | E | 5699 | 1292 | |
| 7 | 所有售货机的订单总量 | | | |
| 8 | 70679 | | | |
| 9 | 所有售货机的交易总额 | | | |
| 10 | 286979.7 | | | |
| 11 | | | | |
| 12 | | | | |

售货机的交易额和订单量

售货机的订单总量和交易总额

## 1.3 计算每台售货机每月的日均订单量、每单平均交易额

先读取数据，然后利用 to_datatime 将提取出相应售货机的支付时间一列数据从字符串格式转换成 datatime 格式，再转为索引。

创建两个空 list，分别存入各个售货机的每月每单平均交易额数据和每月日均订单量数据。

最后将数据储存到 dataframe 格式的 data1、data2 里，最后输出到相应的 csv 文件。

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | A | B | C | D | E |
| 2 | 1 | 4.506567 | 3.753005 | 4.328496 | 3.692664 | 4.680226 |
| 3 | 2 | 3.864035 | 3.255676 | 3.826087 | 3.088652 | 3.638372 |
| 4 | 3 | 3.58549 | 3.614717 | 3.769962 | 4.305729 | 4.305714 |
| 5 | 4 | 4.036913 | 4.07529 | 4.403678 | 3.790293 | 4.159888 |
| 6 | 5 | 4.477646 | 4.236133 | 4.726743 | 4.241312 | 4.410991 |
| 7 | 6 | 4.047394 | 4.06805 | 4.5017 | 4.025962 | 3.817856 |
| 8 | 7 | 4.097689 | 4.401739 | 3.988351 | 4.229653 | 3.919311 |
| 9 | 8 | 3.358709 | 3.5842 | 3.913582 | 3.316503 | 3.804471 |
| 10 | 9 | 4.307212 | 4.130258 | 4.427294 | 3.89939 | 4.125375 |
| 11 | 10 | 4.020703 | 4.11234 | 4.27333 | 3.884233 | 3.676125 |
| 12 | 11 | 4.471552 | 4.268784 | 4.352393 | 3.862314 | 4.283227 |
| 13 | 12 | 3.787868 | 3.667014 | 3.943043 | 3.57258 | 4.168973 |
| 14 | | | | | | |

每月每单平均交易额

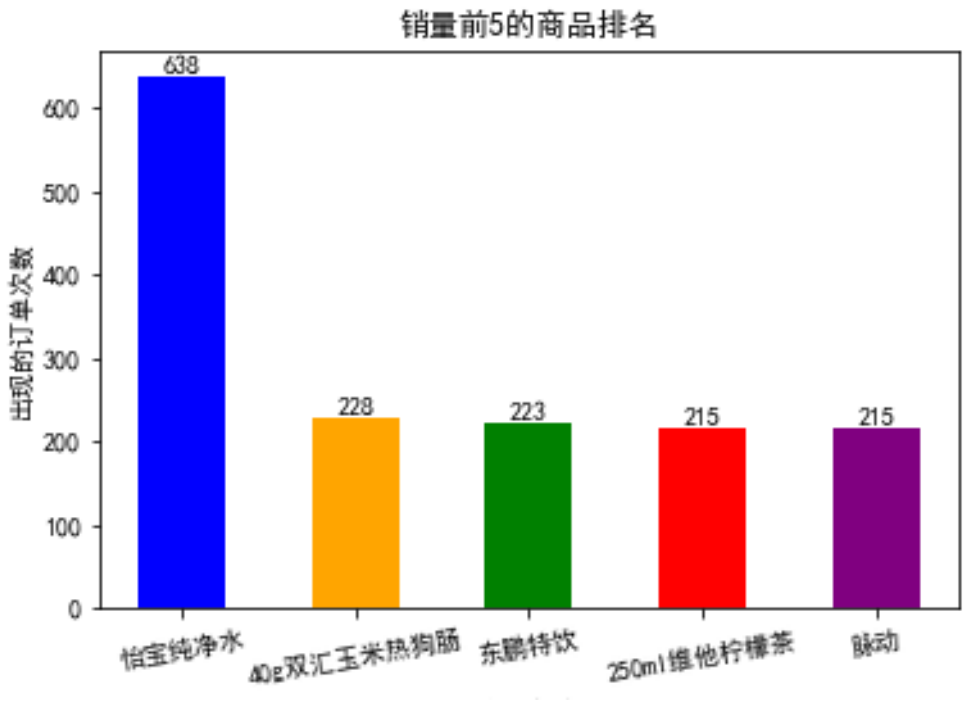
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | A | B | C | D | E |
| 2 | 1 | 10.80645 | 11.80645 | 12.22581 | 8.354839 | 11.41935 |
| 3 | 2 | 4.071429 | 6.607143 | 7.392857 | 5.035714 | 9.214286 |
| 4 | 3 | 8.225806 | 8.548387 | 8.483871 | 6.193548 | 11.29032 |
| 5 | 4 | 14.9 | 20.1 | 24.46667 | 14.76667 | 29.83333 |
| 6 | 5 | 24.3871 | 28.03226 | 25.45161 | 18.19355 | 41.67742 |
| 7 | 6 | 55.63333 | 61.86667 | 62.73333 | 34.66667 | 86.43333 |
| 8 | 7 | 15.35484 | 11.12903 | 24.64516 | 10.22581 | 26.22581 |
| 9 | 8 | 21.48387 | 31.64516 | 40.6129 | 23.06452 | 57 |
| 10 | 9 | 34.66667 | 58.16667 | 55.93333 | 32.76667 | 137.8 |
| 11 | 10 | 50.48387 | 65.35484 | 71.48387 | 38.25806 | 89.58065 |
| 12 | 11 | 38.66667 | 67.7 | 64.76667 | 40.33333 | 167.3333 |
| 13 | 12 | 64.6129 | 71.29032 | 76.74194 | 53.64516 | 104.9032 |
| 14 | | | | | | |

每月的日均订单量

# 任务 2

## 2.1 绘制 2017 年 6 月销量前 5 的商品销量柱状图

先读取数据，如任务 1.2 做法将支付时间一列的数据转换为时间序列数据格式，取出 2017 年 6 月的数据，利用 iloc 切片的方法初步取出商品销量前五的数据。然后利用绘制柱状图的方法直接绘图即可。




由图可知，怡宝纯净水的销量是最高的，远大于其他商品的销量，是销量第二的 40g 双汇玉米热狗肠的 2.8 倍，可以初步考虑增多纯净水或者矿泉水的投放。

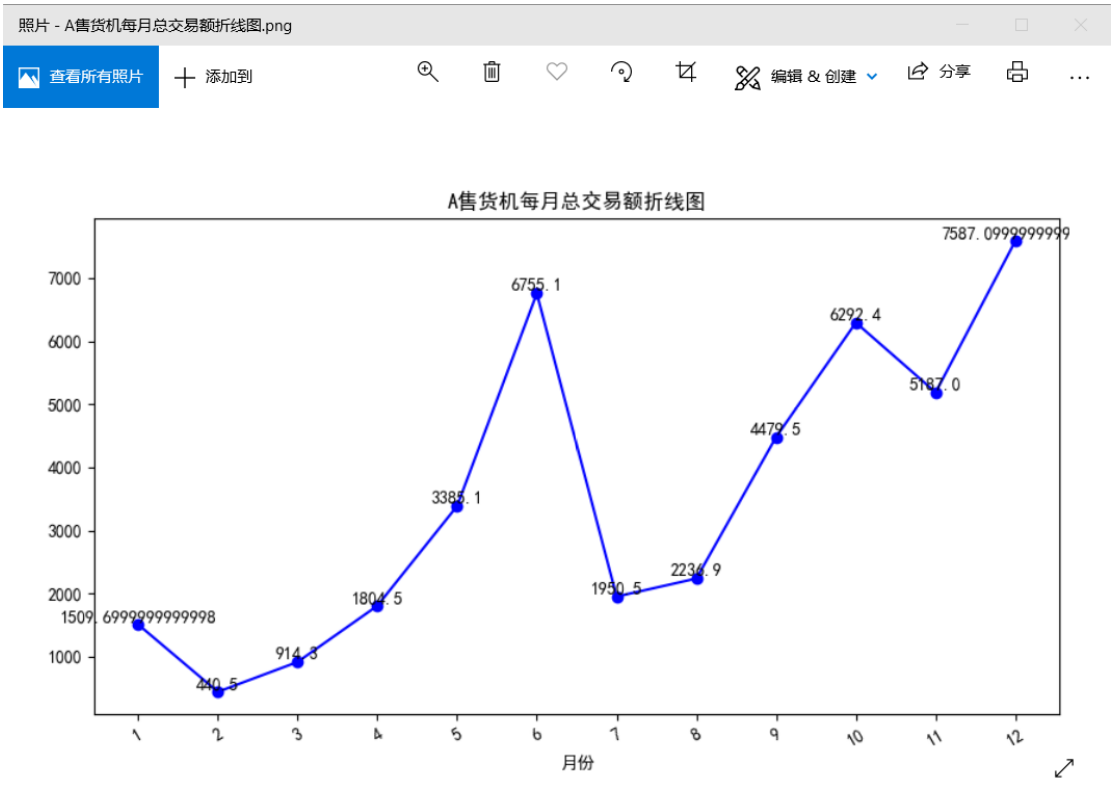
## 2.2

## 每台售货机每月总交易额折线图

先取出数据，设置画布，逐一添加子图。

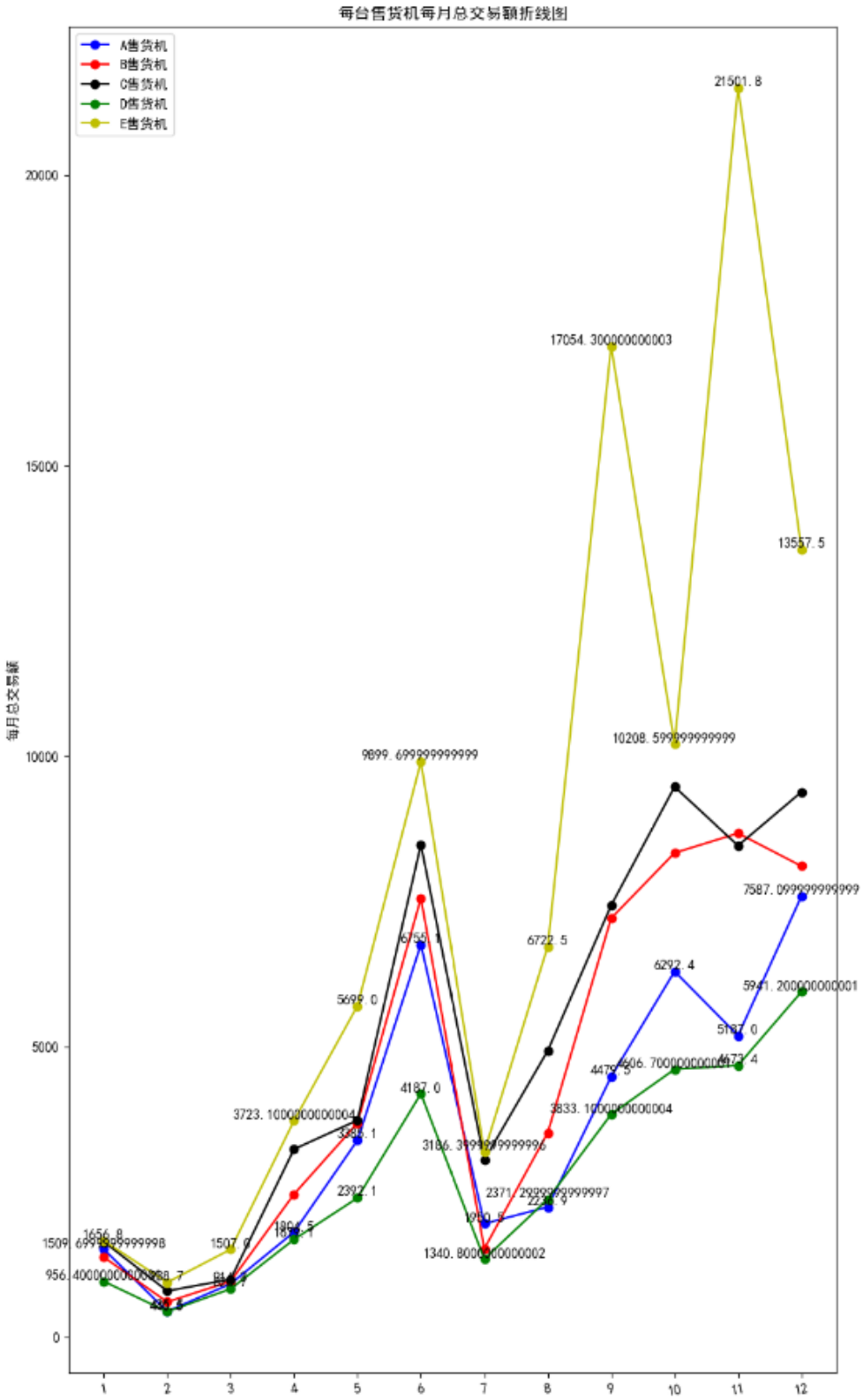
将支付时间一列的数据转换为时间序列数据格式，利用 for 循环逐月取出每台售货机的每月总交易额的数据，利用绘制折线图的方法直接绘图即可。

为了更清楚地了解每台售货机之间的关系，在此任务中，绘制了每幅图分开的情形以及将每幅图放在同一图中的情形。




以 A 售货机为例，可以看出，在前三个月（冬天），由于气候原因，大家不喜欢购买零食饮料，我们可以理解为天气太冷。而从 3 月到 6 月有一个明显的增长，此时属于夏季，大家对饮料零食的需求剧增。

6 到 7 月的交易额暴跌，可以理解为相比夏日炎炎，秋日的凉爽对交易额有影响。而还有一种设想，就是这些售货机都处于校园附近，7、8 月份的交易额低迷可以理解为，进入暑假，学生们都回家了，主力消费者短暂流失。9 月份开学了，交易额又有一个明显的增长，但依旧低于夏日炎炎（6 月）的情形。这都符合数据。难以理解的是不知道为什么 12 月份的数据会处于一个高点，或许还有其他非季节因素对交易额有所影响。



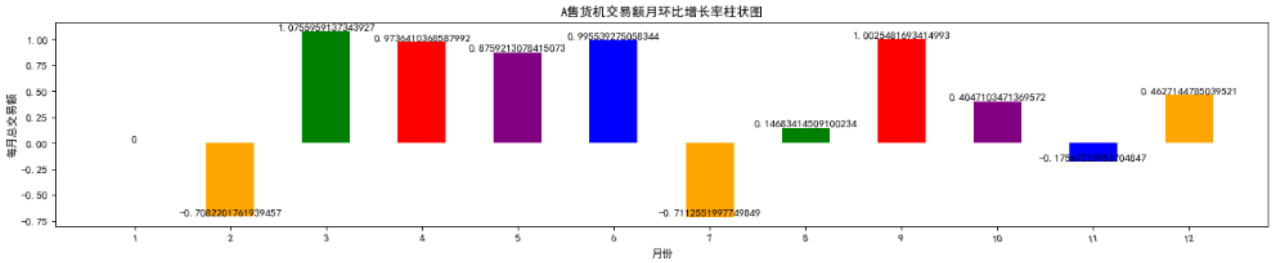

上面是五台售货机的交易额折线图，交易额越高，毛利润就越高。可以看出 E 售货机的交易额比前四台的交易额要明显高许多，可以考虑在 E 售货机附近再增添一台售货机。

D 售货机的交易额明显低于其他的售货机，可以考虑撤掉 D 售货机，改在 E 售货机附近；当然还有一种可能就是 D、E 天然有竞争关系（离得很近），那么可看出 E 的竞争力比 D 高许多，可以考虑将 D 撤掉，换到其他地方，此地区的客户需求由 E 售货机满足即可。

## 交易额月环比增长率柱状图

在绘制了折线图的基础上，创建新 list，根据月环比增长率的定义，利用 for 循环得到每个月的月环比增长率数据，添加到 list 内。然后利用绘制柱状图的方法直接绘图即可。

此处定义每台售货机第一个月的月环比增长率为 0。




（A 售货机的交易额月环比增长率）

以 A 售货机的月环比增长率为例（其余四台的趋势十分相似），可以考虑修改售货机在 2 月、7 月、11 月的商品供应，因为这三个月的交易额有明显的下降，或许是商品供应不符合季节气候情况。

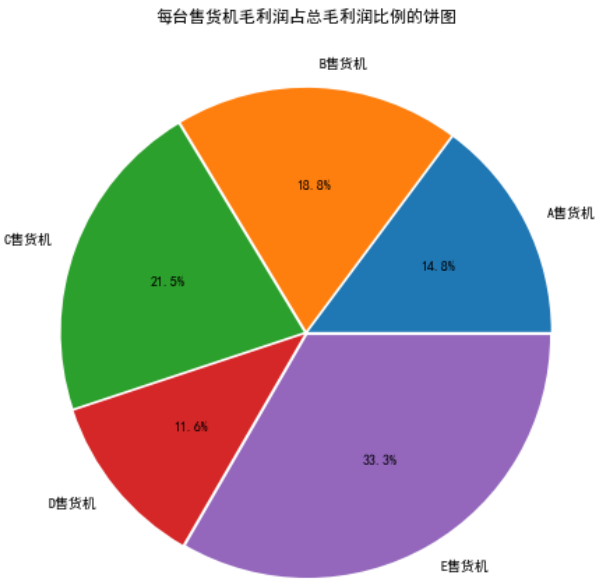
## 2.3 绘制每台售货机毛利润占总毛利润比例的饼图

先读取附件 1（也包括 A 至 E 的数据）、附件 2 数据，利用 groupby 方法将附件 1 中的商品和实际金额列分离出来，对实际金额列求和。

再对附件 2 的数据重新定义索引为商品一列。

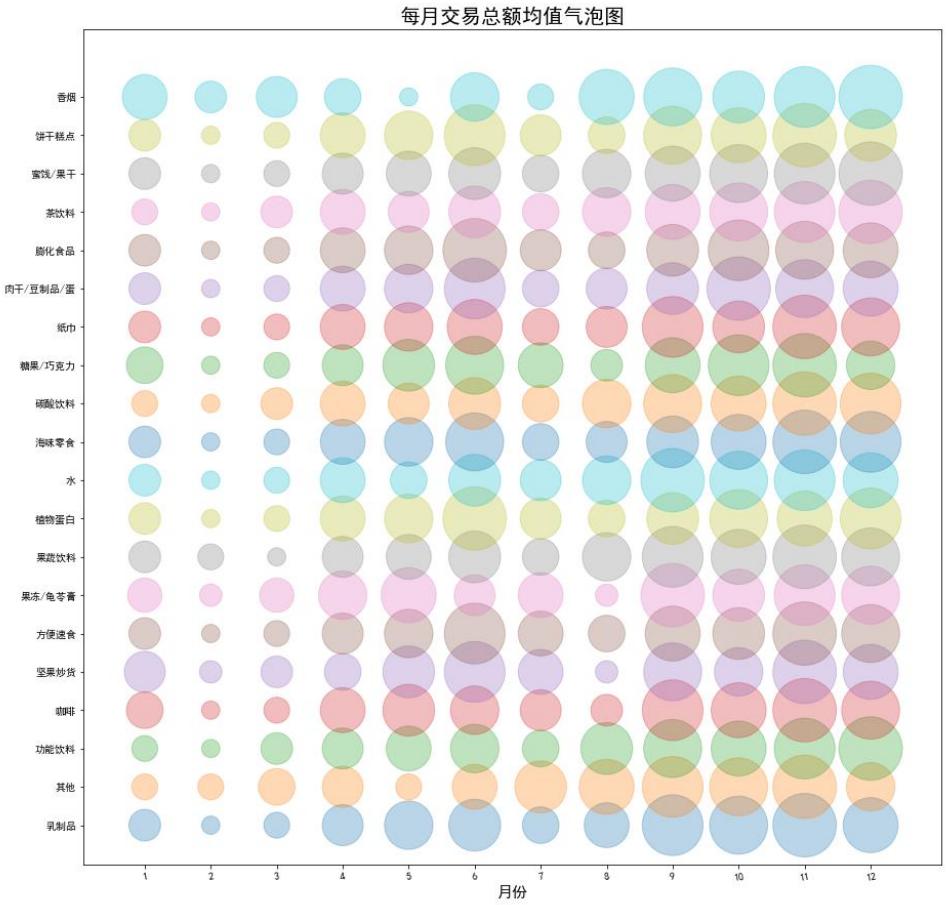
根据索引相同的原理，为附件 1 中的数据贴上饮料/非饮料类的标签，再根据标签利用 if 语句对毛利润进行加权计算，得到每台售货机的毛利润数据。

再利用绘制饼图的语句直接绘制饼图即可。



## 2.4 绘制每月交易额均值气泡图

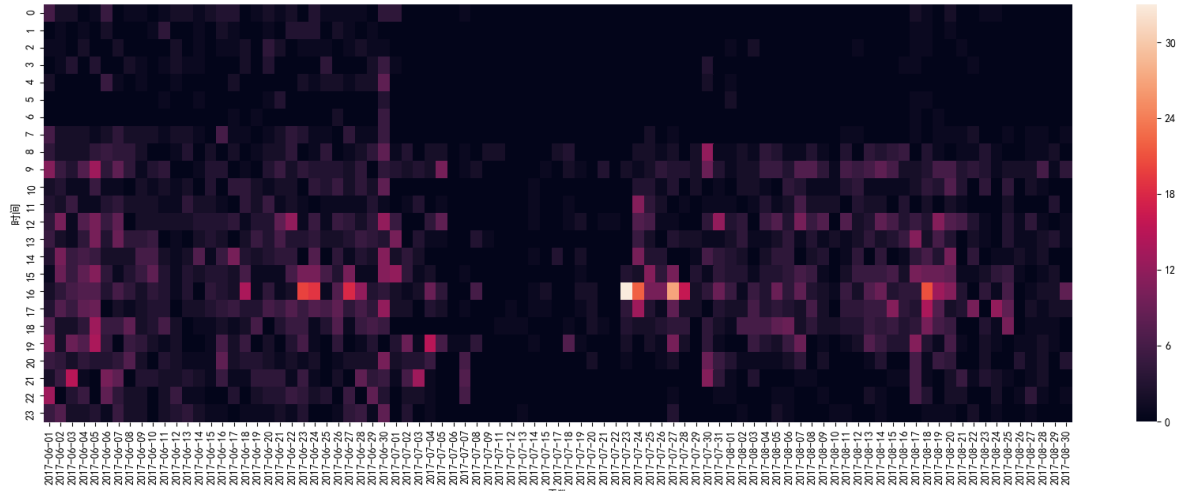
先读取附件 1、附件 2 的数据，将附件 1、2 的数据直接合并在一起转换时间序列数据格式，根据二级类一列求和，再将数据存入 dataframe 格式里。
再利用 for 循环得到实际的每月交易额，根据绘制气泡图的方法绘制气泡图即可。




## 2.5 绘制 C 售货机 678 月订单热力图

先取出数据，将支付时间一列的数据转换为时间序列数据格式，根据需求将支付时间限制在 6、7、8 月。再利用 groupby 方法将数据分组提取，转换数据为 array 格式存入字典，再将字典转入 dataframe 格式数据，再利用 drop_duplicates 去重，利用根据列对数据表进行重塑，利用 fillna 处理缺失值部分（补充为 0），最后利用 heatmap 方法绘制热力图即可。

如下热力图，颜色越浅代表订单量越多，反之则是订单量少的情形。可以发现，一般来讲，早上九点之后就陆陆续续有消费者购买商品；而每日的高峰在每天的下午三点到五点之间。一个合理的解释应该是这个时间段大家都出来运动，需要补充能量、饮料。




678 月订单热力图

# 任务 3

## 3.1 为各个售货机的商品贴标签

先取出各个售货机的数据以及，附件 2 的数据，利用 merge 方法直接合并 dataframe。再利用 isin 方法取出饮料类数据。再利用 for 循环以及 if 语句判断商品的标签，并将生成的 dataframe 储存到相应 csv 文件中。

此处定义滞销商品为年销量小于 24 份，热销商品定义为年销量大于 180 份。

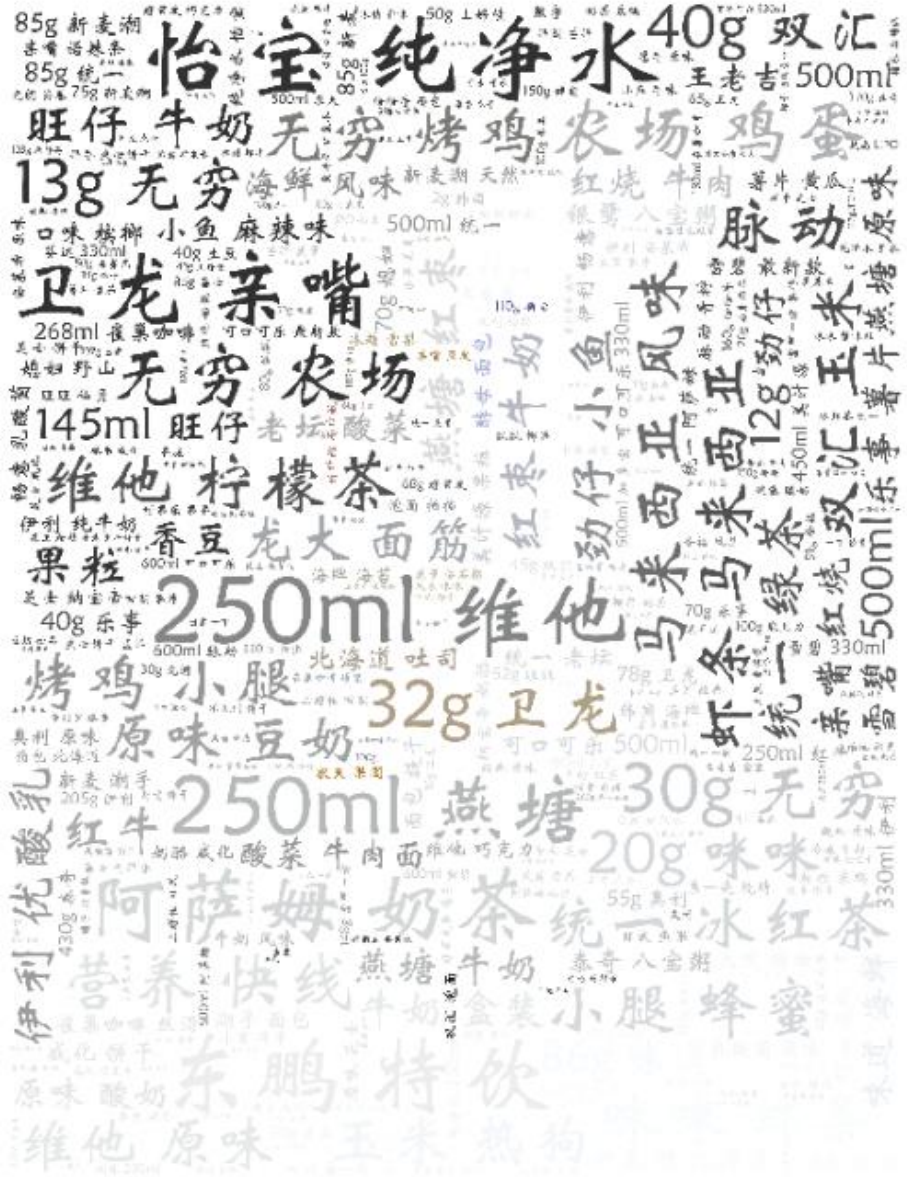
| | 商品名 | 订单量 | 标签 |
|---|---|---|---|
| 0 | 100g*5瓶益力多 | 40 | 正常 |
| 1 | 13g雀巢咖啡1+2特浓 | 4 | 滞销 |
| 2 | 145ml旺仔牛奶盒装 | 131 | 正常 |
| 3 | 145ml旺仔牛奶罐装 | 2 | 滞销 |
| 4 | 150g健能酸奶原味 | 28 | 正常 |
| 5 | 180ml雀巢咖啡罐装 | 22 | 滞销 |
| 6 | 205g伊利安慕希原味 | 53 | 正常 |
| 7 | 205ml安慕希蓝莓味 | 7 | 滞销 |
| 8 | 250ML东鹏特饮 | 16 | 滞销 |
| 9 | 250ml燕塘原味酸奶 | 112 | 正常 |
| 10 | 250ml燕塘甜牛奶 | 126 | 正常 |
| 11 | 250ml燕塘红枣牛奶 | 128 | 正常 |
| 12 | 250ml王老吉盒装 | 22 | 滞销 |
| 13 | 250ml红牛 | 71 | 正常 |
| 14 | 250ml统一麦香奶茶 | 16 | 滞销 |
| 15 | 250ml维他原味豆奶 | 176 | 正常 |
| 16 | 250ml维他奶低糖原味 | 25 | 正常 |
| 17 | 250ml维他奶巧克力味 | 69 | 正常 |
| 18 | 250ml维他奶黑豆奶饮品 | 11 | 滞销 |
| 19 | 250ml维他柠檬茶 | 181 | 热销 |
| 20 | 250ml维他椰子植物蛋白饮 | 8 | 滞销 |
| 21 | 250ml维他椰子植物蛋白饮 | 6 | 滞销 |
| 22 | 250ml香满楼纯牛奶 | 2 | 滞销 |
| 23 | 268ml雀巢咖啡丝滑拿铁 | 91 | 正常 |
| 24 | 330ml伊利畅意乳酸菌原味 | 74 | 正常 |
| 25 | 450ml美汁源果粒橙 | 69 | 正常 |
| 26 | 480ml小茗同学冷泡溜溜哒 | 7 | 滞销 |
| 27 | 480ml小茗同学冷泡青柠红 | 36 | 正常 |
| 28 | 500ML宝矿力 | 4 | 滞销 |

task3_1A

（示例：为 A 售货机饮料贴标签）

## 3.2 绘制各个售货机的画像（词云图）

将 A、B、C、D、E 售货机的商品名称、支付时间数据拷贝另存到相应的词云图来源文本中，制作相应的词云图分析




A 售货机词云图

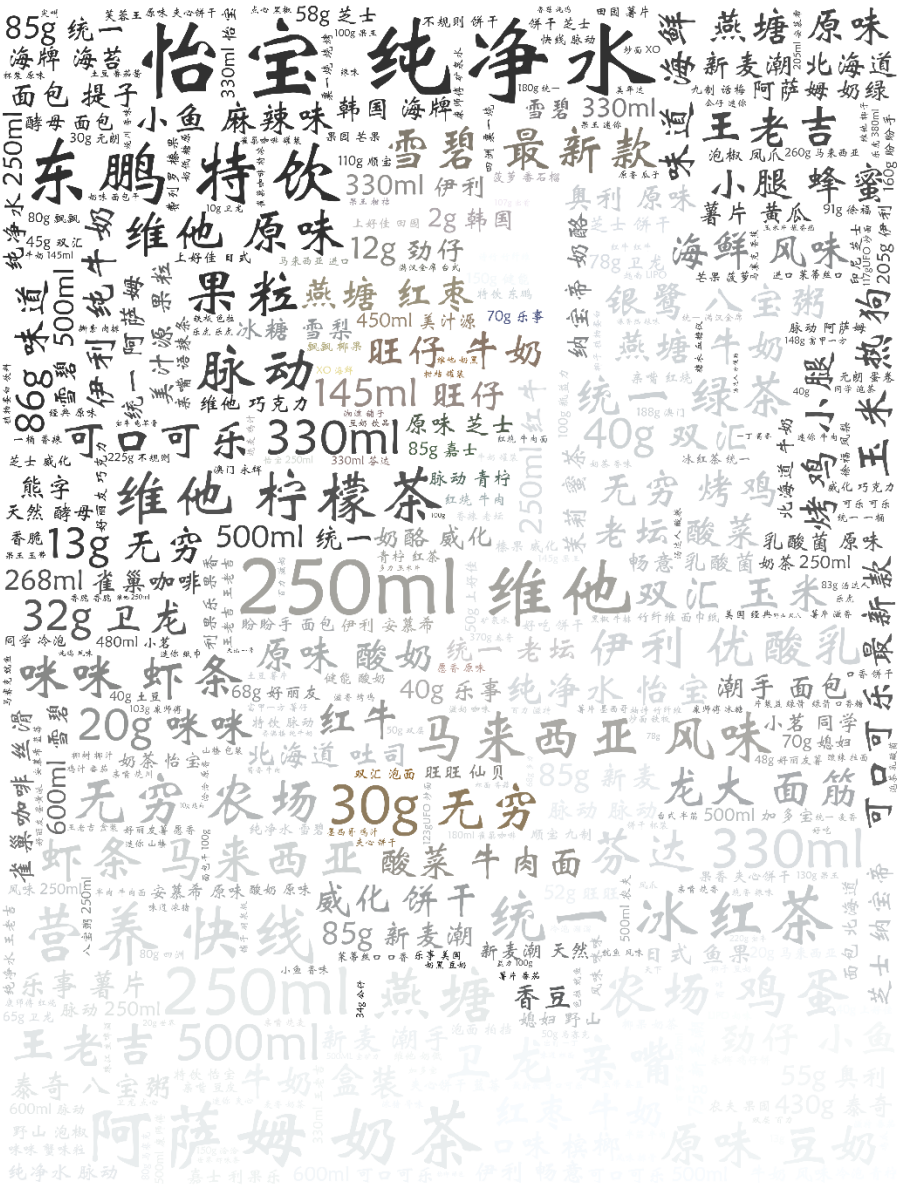
可以看出，怡宝纯净水是最好卖的商品，A 售货机可以考虑增加纯净水/矿泉水的商品位置；

250ml 也是很明显的一个大词，说明这个容量的饮料更受消费者喜爱，添加饮料建议添加 250ml 饮料的位置

还有明显的卫龙亲嘴烧、无穷、维他柠檬茶、马来西亚风味、阿萨姆奶茶、东鹏特饮都是热词，可以添加与此相关的商品。


B 售货机的词云图

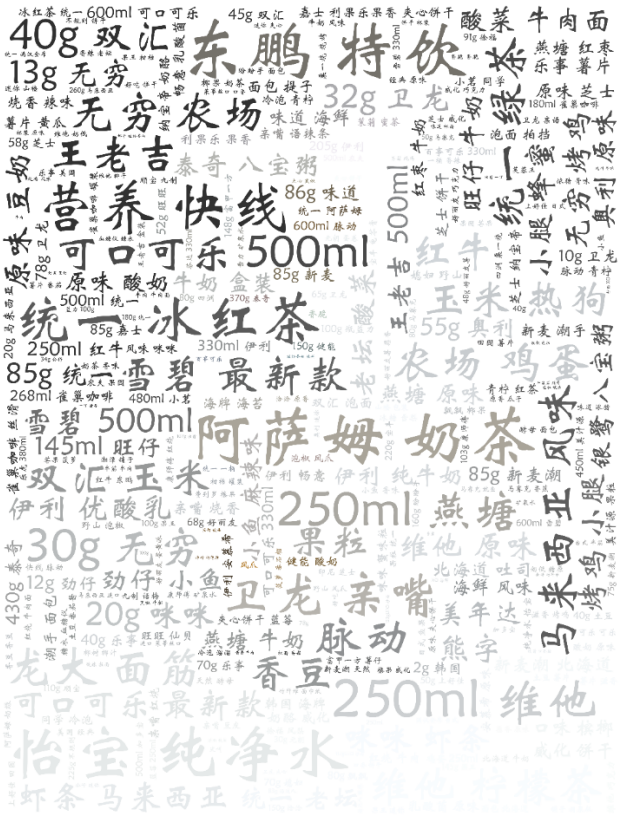
可以看出，怡宝纯净水也是最好卖的商品，B 售货机也可以考虑增加纯净水/矿泉水的商品位置；还有非常显眼的东鹏特饮、可口可乐、阿萨姆奶茶、营养快线。



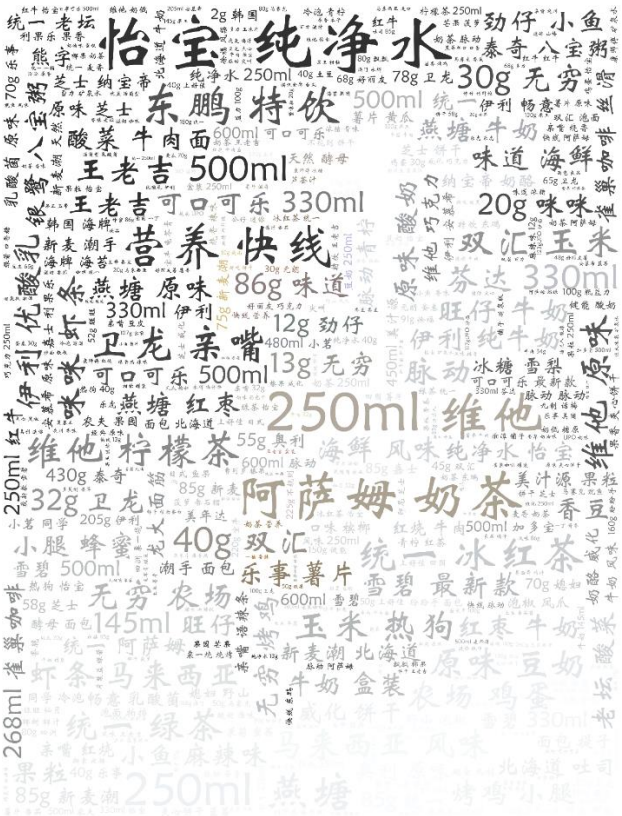

C 售货机的词云图

如同 A、B 售货机，C 词云图最明显的最好卖的也是怡宝纯净水；不一样的是出现了一个“最新款”的字眼，说明新款对消费者是有一定吸引力的。还有非常显眼的统一冰红茶、阿萨姆奶茶、营养快线，这些饮品可以在这台售货机再增加位置。

对 D、E 售货机可以做类似的分析。




D 售货机的词云图




E 售货机的词云图

# 任务 4

## 4.1

预测未来销售额的原理是利用已知的销售额的完整数据，对此进行数据清洗和预处理，分析和建模，通过如灰色预测方法，时间序列分析方法等，找出描绘数据变化的内在规律或者说是方程，从而对未来的销售额进行预测。

我认为附件所提供的数据不足以对 2018 年 1 月每台售货机的每个大类商品的交易额进行预测。因为对于这组数据，可以依据基本的常识，以及任务 2.2 的 A 至 E 五台售货机的折线图可以看出其交易额会受到季节性因素（周期性）的影响。而所给数据只涵盖了一年的数据，相当于只有一个周期的数据，无法根据一个周期的数据分析预测出数据的走势，至少需要两个周期的数据。

## 4.2

如果项目的需求方可以提供两年及两年以上的售货机销售数据，那么便可以通过灰色预测方法得到 2018 年 1 月的交易额的预测交易额。数据量越大，那么理论上的预测会越加精确。