目录

一、 背景介绍

1、 行业背景

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、自助的经营模式节首人工成本，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。

2、 分析目标

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 提供了商品的分类情况。

其一，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。

其二，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。

3、 数据说明

数据来源：某地 ABCDE 五个自动售货机销售记录

数据起止时间：2017 年 1 月 1 日——12 月 31 日

销售信息数据量：70680 条记录（其中有一条为无效数据）

货品类别数据量：315 条记录

二、 任务一 数据预处理与分析

1、 数据预处理

通过 Python 自带的 pandas 模块对附件 1 进行数据预处理，得到 ABCDE 五个自动售货机在售所有货品的价格、数量、属性、分类。

销售信息表的长度为： 70679
销售信息表为：

	订单号	设备ID	应付金额	实际金额	商品
0	DD20170816749366361849990978 4	E43A6E078A07631	4.5	4.5	68g好丽友巧克力派2枚
1	DD20170816749366355581406116 4	E43A6E078A04172	3.0	3.0	40g双汇玉米热狗肠
2	DD20170816749357852689093988 6	E43A6E078A06874	5.5	5.5	430g泰奇八宝粥
3	DD20170816749368350718661583 7	E43A6E078A04228	5.0	5.0	48g好丽友薯愿香烤原味
4	DD20170816749375954861825200 6	E43A6E078A04134	3.0	3.0	600ml可口可乐

	支付时间	地点	状态	提现
0	2017/1/1 0:53:00	D	已出货未退款	已提现
1	2017/1/1 1:33:00	A	已出货未退款	已提现
2	2017/1/1 8:45:00	E	已出货未退款	已提现
3	2017/1/1 9:05:00	C	已出货未退款	已提现
4	2017/1/1 9:41:00	B	已出货未退款	已提现

表 1 销售信息表导入

货品类别表的长度为： 315
货品类别表为：

	商品	大类	二级类
0	100g*5瓶益力多	饮料	乳制品
1	100g越南LIPO奶味面包干	非饮料	饼干糕点
2	10g卫龙亲嘴烧香辣味	非饮料	肉干/豆制品/蛋
3	10g越南LIPO奶味面包干	非饮料	饼干糕点
4	110g顺宝九制话梅	非饮料	蜜饯/果干

表 2 货品信息表导入

2、 任务 1.2

通过图表可知，2017 年 5 月五台售货机的订单量均值均在 4 左右分布。其中，D 售货机的订单总量和交易总额最低，分别为 564 和 2392.1；E 售货机的订单总量和交易总额最高，分别为 1291 和 5696。通过全年统计，五台售货机订单总量为 70679，交易总额为 286979.7。

	A	B	C	D	E	总计（全年）
mean	4.477646	4.219815	4.720051	4.241312	4.412084	4.060325
std	3.909918	3.429409	4.375095	3.146491	3.021968	3.357931
min	0	0	0	0	0	0
25%	3	3	3	2.5	3	3
50%	4	3.5	4	3.5	3.5	3.5
75%	5	5	5.5	4.5	5.5	4.5
max	51	72	72	25	54.5	125
订单总量	756	863	788	564	1291	70679
交易总额	3385.1	3641.7	3719.4	2392.1	5696	286979.7

表 3 每台售货机 5 月份交易额、订单量等信息汇总

3、 任务 1.3

从图 3 可以看出，五台自动售货机的每单平均交易额均在 3 到 4 左右，呈现较为平稳的

分布。 纵向比较来看，ABC 三台自动售货机在夏季和秋季月份间平均交易额较大，D 自动售货机的货品平均交易额主要在夏季增长，而 E 售货机在春秋冬季的平均交易额均在 4 以上。横向比较来看，夏季的平均交易额大体在 4 以上，售货机的盈利效果较好。

在日均订单量方面，五台自动售货机在冬季月份的订单量明显大于其他月份，而春季的订单量最少。导致这种现象的原因可能是春节期间过年在家人口数变多，五台自动售货机的使用频率增加，从而使得订单量增加。春季人口回归，使得五台自动售货机的日均订单量有一个陡崖式的下降。

月\类	A	B	C	D	E
1	4.507	3.753	4.328	3.693	4.680
2	3.864	3.256	3.827	3.089	3.638
3	3.585	3.615	3.770	4.306	4.306
4	4.037	4.075	4.404	3.790	4.160
5	4.478	4.236	4.727	4.241	4.411
6	4.047	4.068	4.502	4.026	3.818
7	4.098	4.402	3.988	4.230	3.919
8	3.359	3.584	3.914	3.317	3.804
9	4.307	4.130	4.427	3.899	4.125
10	4.021	4.112	4.273	3.884	3.676
11	4.472	4.269	4.352	3.862	4.283
12	3.788	3.667	3.943	3.573	4.169

表 3 每台售货机每月每单平均交易额

月\类	A	B	C	D	E
1	10.806	11.806	12.226	8.355	11.419
2	4.071	6.607	7.429	5.036	9.214
3	8.226	8.548	8.484	6.194	11.290
4	14.900	20.100	24.467	14.767	29.833
5	24.387	28.032	25.452	18.194	41.677
6	55.633	61.867	62.733	34.667	86.433
7	15.355	11.129	24.645	10.226	26.226
8	21.484	31.645	40.613	23.065	57.000
9	34.667	58.167	55.933	32.767	137.800
10	50.484	65.355	71.484	38.258	89.581
11	38.667	67.700	64.767	40.333	167.333
12	64.613	71.290	76.742	53.645	104.903

图 4 每台售货机每月日均订单量

三、 任务二 数据分析与可视化

1、任务 2.1

由图 5 可以看出，2017 年 6 月销量前五的商品有怡宝纯净水、脉动、东鹏特饮、250ml

维他柠檬茶和营养快线。其中怡宝纯净水销量最高，需求量最大，远远领先其他四种商品，为 1754 元，营养快线销量最低，为 764.5 元。

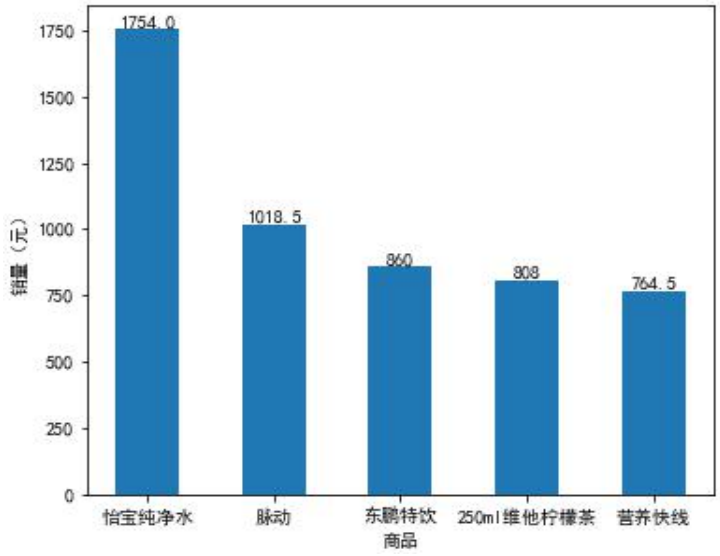


图 5 6 月销量前五的商品

2、任务 2.2

五台售货机在 2017 年 12 个月份的交易额如下图所示。从整体来看，五台售货机总交易额较大的波动均出现在夏季和冬季。春季的总交易额在全年中较低，且随着月份增加而不断增长，在 6 月份达到峰值。其后，五台售货机的总交易额均在 7 月份有一个明显的降低趋势，再随着冬季的到来而上升。

对于 BCD 三台售货机而言，在春夏秋三季的总交易额波动情况较为相似，A 和 E 售货机在 10 到 12 月份的时候经历一个下降的波动，尤其 E 售货机的波动较为剧烈，9 月和 11 月有两个极大峰值，10 月和 12 月有两个极小峰值。从总交易额的大小来看，E 售货机的交易额最大，11 月份达到 21501.8 元。D 售货机总交易额最小。

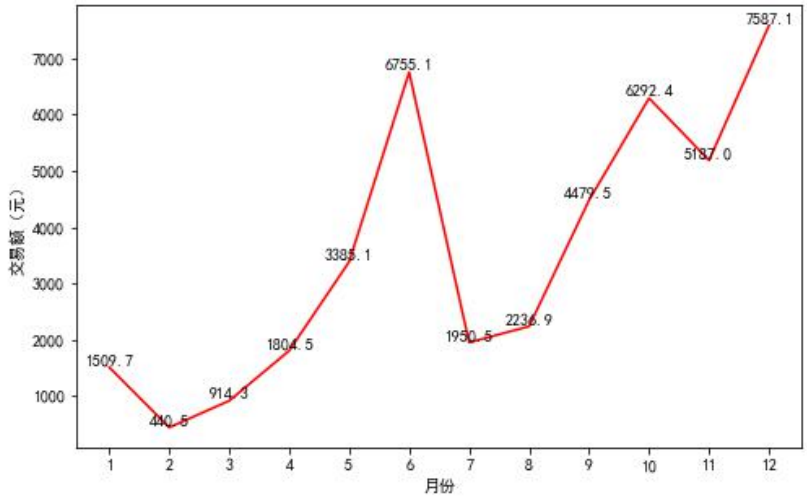


图 6 A 售货机各月总交易额

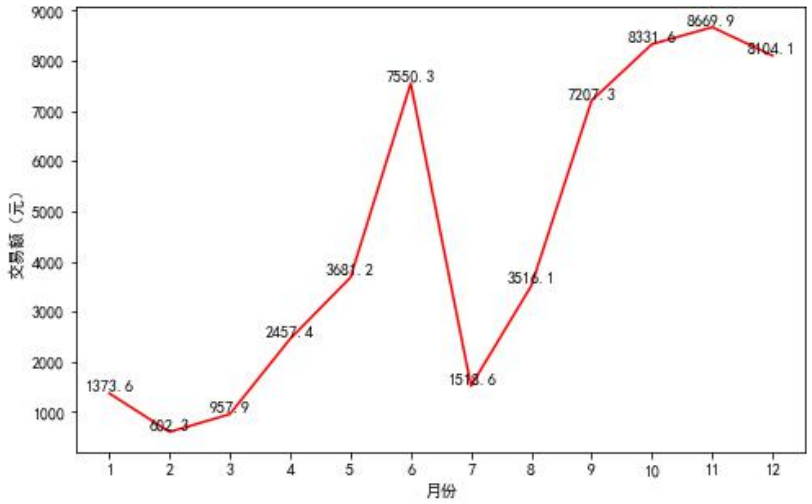


图 7 B 售货机各月总交易额

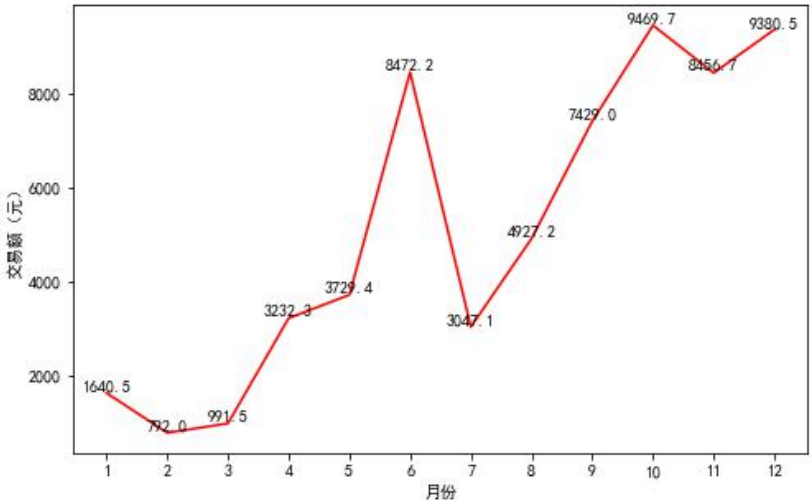


图 8 C 售货机各月总交易额

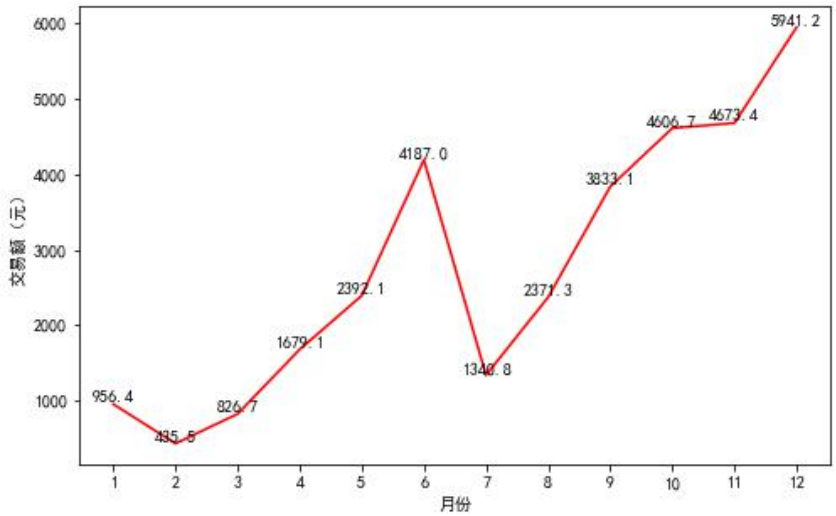


图 9 D 售货机各月总交易额

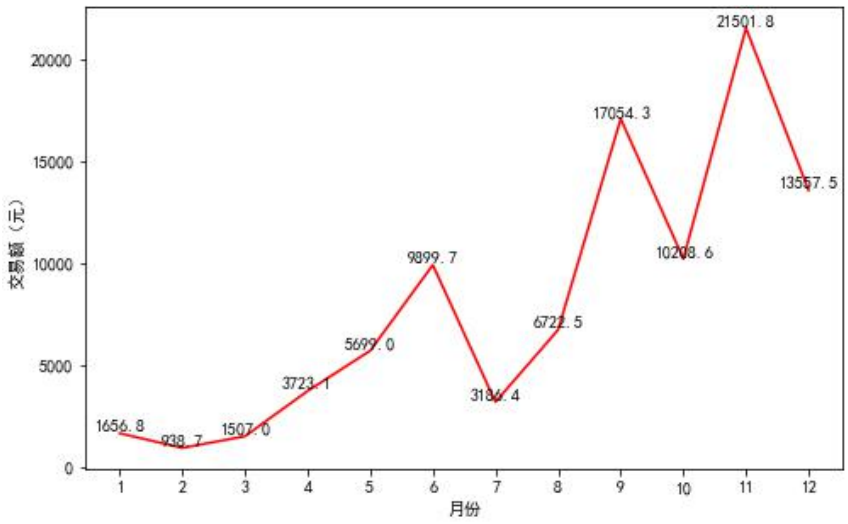


图 10 E 售货机各月总交易额

通过观察五台售货机交易额的月环比增长率可以发现，3 月和 4 月的环比增长率是全年中最高的，其次为 9 月份。环比增长率在 2 月和 7 月均为负值，说明总交易额出现下降的现象。类比学校来说，这两个月正值寒暑假期间，售货机的使用率较平时小。

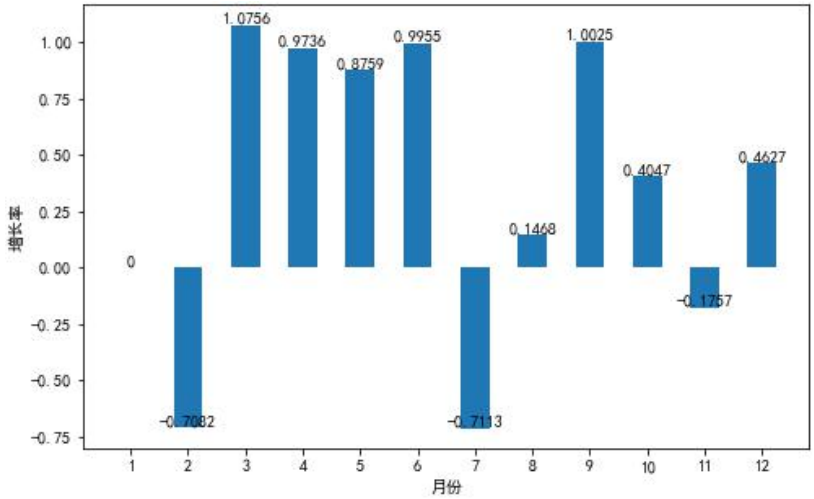


图 11 A 售货机交易额月环比增长率

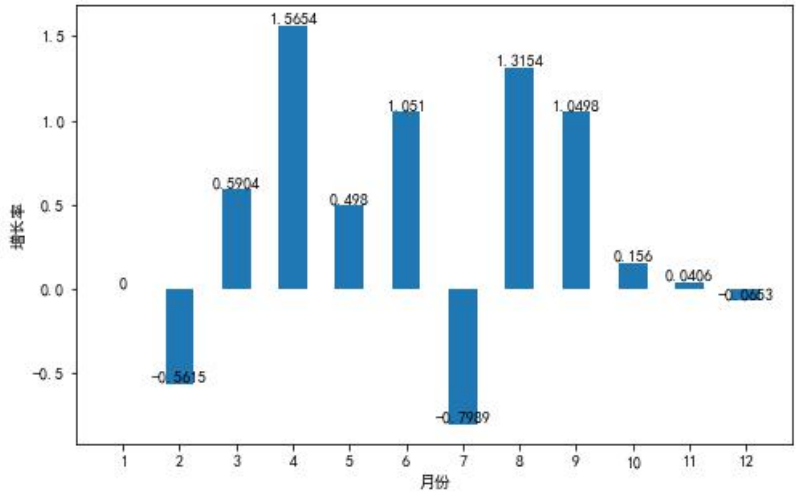


图 12 B 售货机交易额月环比增长率

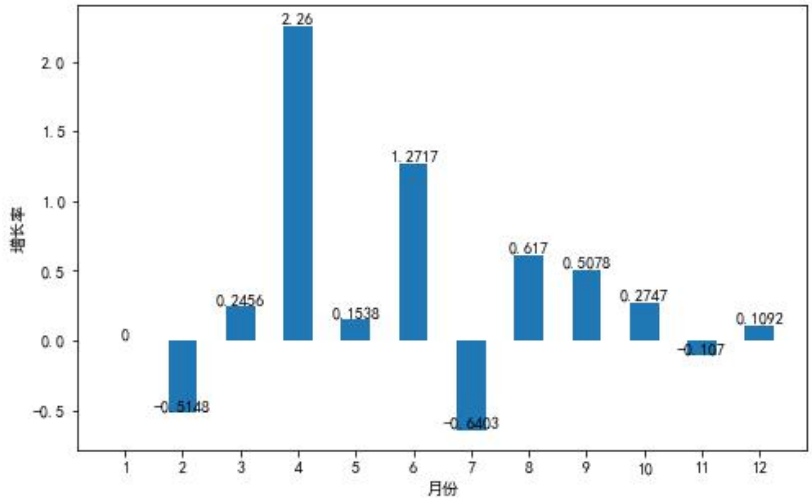


图 13 C 售货机交易额月环比增长率

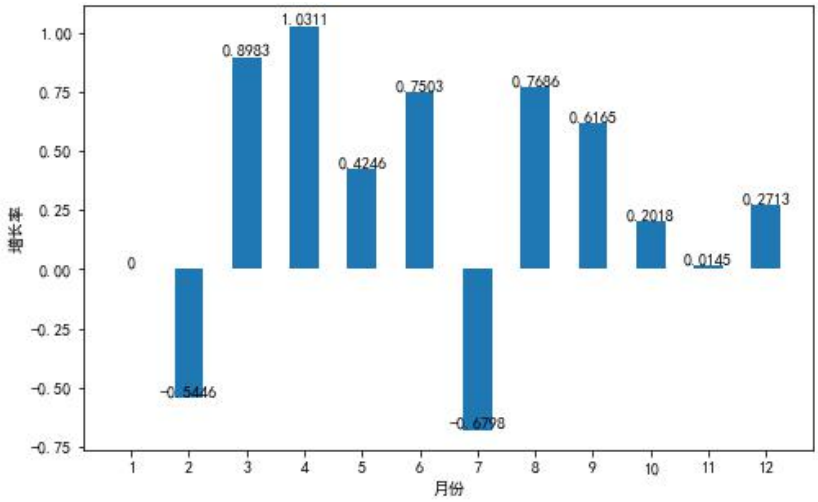


图 14 D 售货机交易额月环比增长率

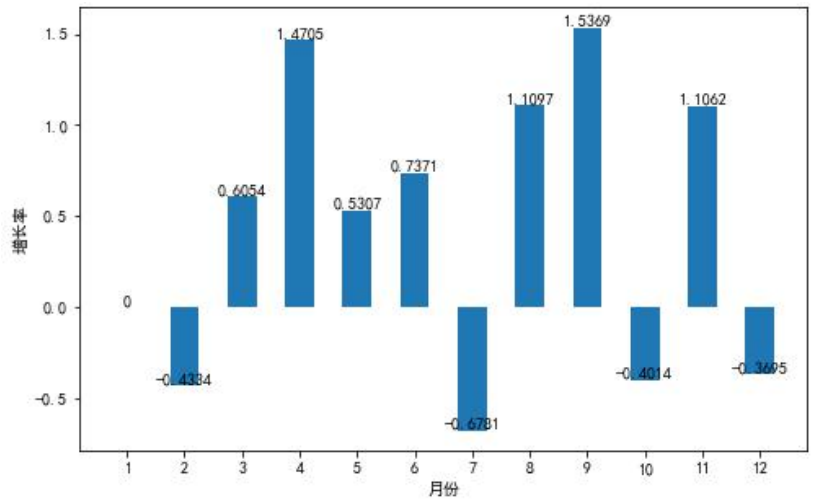


图 15 E 售货机交易额月环比增长率

3、任务 2.3

通过计算得到饼图如下，毛利润比例由大到小依次为 ACBED。说明 A 的实际盈利价值要优于其他四台售货机，饮料类的销售较好。对于 D 售货机，非饮料的销售比例较大些。

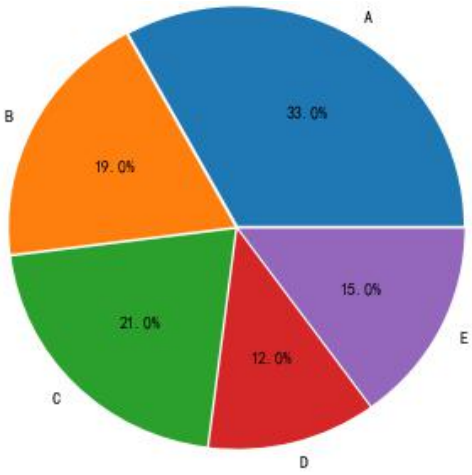


图 16 各售货机毛利润占总毛利润比例

4、任务 2.4

根据商品所属二级类，画出横轴为月份，纵轴为二级类标签的交易额均值气泡图如下。其中可以看到香烟的交易额均值最大，水的交易额均值最小，符合实际生活中的定价。特别注意的是，有些二级类商品在某几个月没有交易额。

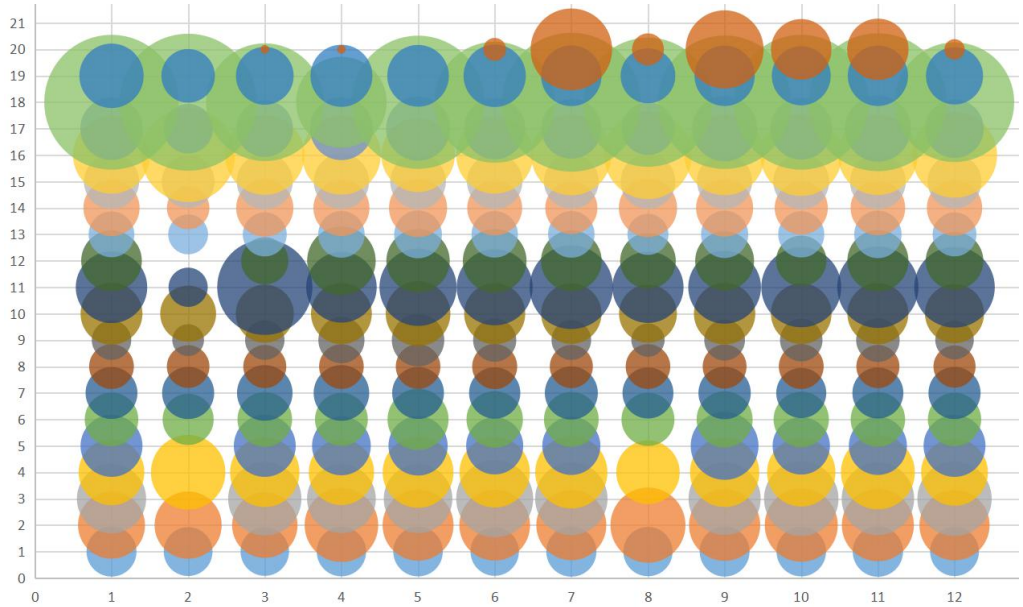


图 17 每月交易额均值气泡图（不同颜色代表不同二级类商品）

碳酸饮料	1	果蔬饮料	6	糖果/巧克力	11	蜜饯/果干	16
咖啡	2	植物蛋白	7	纸巾	12	饼干糕点	17
坚果炒货	3	水	8	肉干/豆制品/蛋	13	香烟	18
方便速食	4	海味零食	9	膨化食品	14	乳制品	19
果冻/龟苓膏	5	功能饮料	10	茶饮料	15	其他	20

图 18 y 坐标对应的商品二级类标签

四、 任务三 生成自动售货机画像

1、 任务 3.1

分析五台售货机的销售数据，总结规律得到，饮料类商品较为热销的是水、柠檬茶、功能型饮料（东鹏特饮）以及牛奶，滞销的大多数是市面上不常见的饮料商品。分类的依据是销量的前 25%为热销商品，后 25%为滞销商品，其余中间 50%为正常销量的商品。

标签详见 task3-1A/B/C/D/E.csv 五个文件。

2、 任务 3.2

通过 python 的词云图画像分析五台售货机的销售数据得到以下的五个画像，可以发现 ABCE 四台售货机销量最高为怡宝纯净水，D 售货机销量最高为东鹏特饮。

根据这些画像的异同特点可以提取出非常具有商业价值的信息，有利于商家进一步挖掘售货机的需求商机。通过对比不同售货机不同销量的产品来及时调整进货补货的量度，该画像以及背后的数据很具有参考意义。

图 20 A 售货机画像

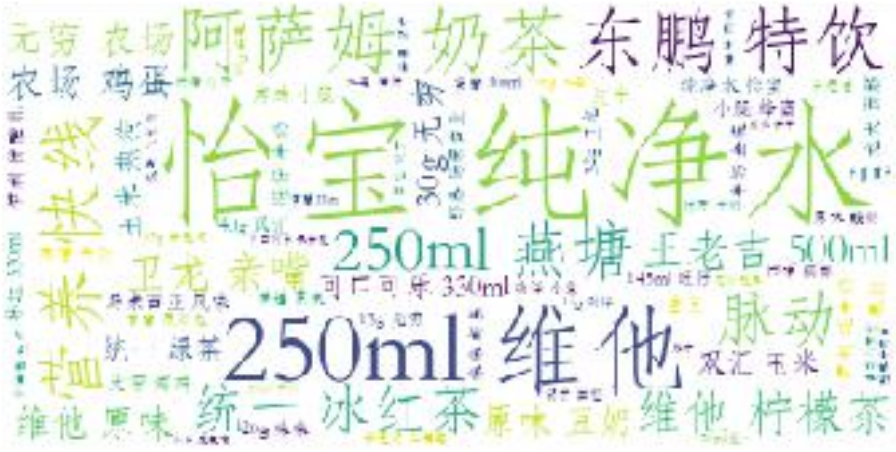


图 21 B 售货机画像

图 22 C 售货机画像

图 23 D 售货机画像

图 24 E 售货机画像

五、任务四 业务预测

1、任务 4.1

预测未来销售额的原理，即机器学习的整体流程为：首先，将数据集划分为训练集和测试集，其次，对于训练集做特征筛选，提取有信息量的特征变量，而筛除掉无信息等干扰特征变量，再次，应用算法建立模型，最后，结合测试集对算法模型的输出参数进行优化。

这里主要介绍集中算法预测模型。有线性回归模型、决策树（回归树）模型、随机森林、xgboost、神经网络、支持向量回归等六种算法模型。

线性回归模型：假设销量与影响销量的因素是线性关系的，包括误差分布、线性方程和激活函数等。通常连续型数值的预测可以用称为回归的统计技术进行建模。回归分析的目的是找到一个联系输入变量和输出变量的最优模型。

决策树（回归）：其原理是通过 if-then 规则对特征变量进行逐步决策来构建的模型。此处，可以举一个例子来简单讲解决策树算法的思想是什么？比如说我想给一个妹纸进行颜值评分，分值范围为[0,10]。评分的第一轮判断是五官是否端正？如果为否，打 3 分；如果为是，则进行第二轮判断，即身材，身材不好则打 5 分。身材好的话再进入第三轮判断，即是否有钱，有钱就是典型的白富美，就是 9 分。没钱则为 7 分。从图中可以看出，其判断决策的过程倒过来看是一个树，红色是它的叶子，叶子对应他的分值，黄色是变量。

随机森林是从决策树演变而成的一个算法，但其思想与决策树相比增加了集成思想。同时，其“随机”具有两层含义，第一层是对特征变量进行随机选择。第二层是，对训练集样本进行随机选择。

xgboost 是基于传统的 GBDT 算法进行了优化的集成算法，它是数据挖掘大赛上面得分非常高的算法。它的思想是这样的，我给一个数据集，我现在有一个问题，就是要看他一家人当中是否会喜欢电子游戏，也是通过构建树的情况进行判断，比如年龄、性别进行判断，它会反映这个家庭成员对应的我们的样本会打一个分，最后男孩给 2 分，女孩给 1 分。有时候我们一棵树确定不了，我们就规定多棵树，树 1 和树 2 之间并不是独立的，第一棵树的时候对样本做第一次判断，判断的时候有对和错，但是我会更关注于我判断错的那一部分，我在规定第 2 棵树的时候，我把预测错的更多的考虑一下，就会变成第 2 棵树，我会过多的关注那些预测错的，再依次的来进行优化。

神经网络是指模拟大脑神经元的工作的非线性模型，神经网络是现在最火的一个深度学习的基础。其包括三个部分：输入层、隐藏层和输出层。输入层在销售预测中则为影响销量变换的各相关因素变量；输出层为销量；中间隐藏层为各相关因素变量到销量之间的一个非线性映射关系，通常为一个函数。

神经网络是在反欺诈领域用得比较多，像现在的银行、互联网金融，有的人进行欺骗性的贷款，就用神经网络可以很快的把他发现出来。还有检测病人也可以用到神经网络。

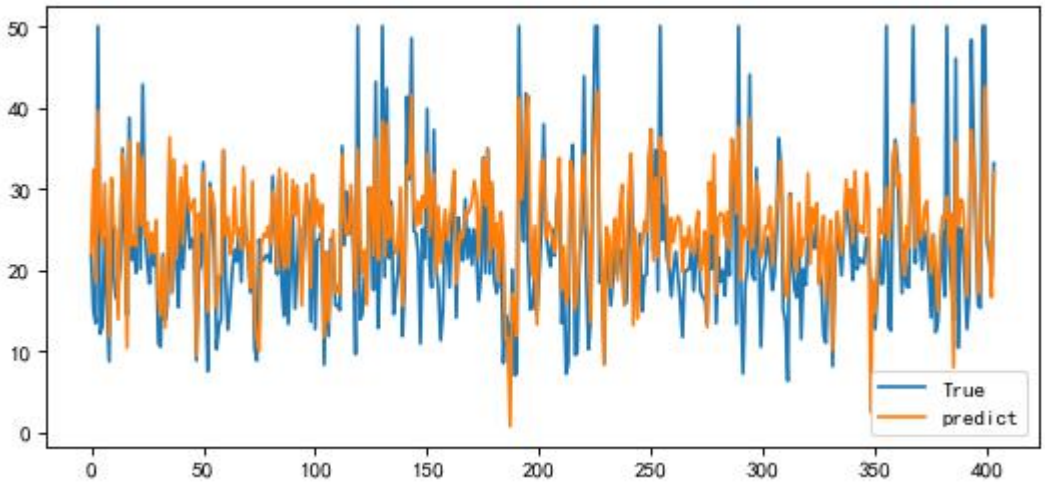
支持向量回归其本质是跟 SVM 是一样，即寻找能使回归局域更大的 margin，其适用于小数据集和高维数据集。

2、任务 4.2

通过以上数据分析，将数据导入 python 中，使用 scikit-learn 中 LinearSVR 方法来构建模型，可以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，且预测值与真实值的波动相差较小，符合周期性规律。

销售预测效果评估中第一方面，方法论 K 折交叉验证。其基本思想为：将总数据集均匀划分为 k 等份(假设取 k=10)，第一次对数据集进行划分过程为：第一份作为测试值，验证这个模型，剩下第 2 到第 10 个做训练集。第二次划分过程为：把第 2 个作为测试值，剩下 9 个作为训练集，然后依次进行训练集和数据集划分，一共会，得到 10 个模型，选择最小的作为我们最终的模型。

第二方面是评估指标 RMSE，值越小，说明预测值与真实值之间的差异就越小，模型效果就越好。


50
40
30
20
10
0
0
50
100
150
200
250
300
350
400
True
predict