新零售无人智能售货机商务数据分析

摘　要

随着科技的迅速发展和智能的广泛应用，便捷、舒适的生活成为人们的追求，自动售货机的发明和应用正迎合了这一要求。因此，运用数据挖掘技术和商务数据分析对自动售货机内商品的供给频率、种类选择、供给量、站点选择等进行分析，帮助经营者掌握市场信息，了解用户需求，制定经营策略和经营方向，对自动售货机这一经营模式的发展有着非常重要的意义。

对于任务一，通过 python 对给出的附件一数据进行重复值、缺失值、异常值的检测和处理，同时删除不合理的数据，得到不重复的自动售货机购买记录。再对附件一中不同地点的售货机数据进行提取，分别保存。之后分别对各个地点售货机总的交易总额和总订单量，以及每台售货机每月的交易额、订单量、每单平均交易额和日均订单量进行统计分析，得出自动售货机的销售情况。

对于任务二，根据任务一得出的结果，利用 python 对其进行数据可视化，根据图像和销售数据的统计分析结果，得出每台售货机的销售规律，利用这些规律，可以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。

对于任务三，分析各个售货机的商品销售数据，观察数据并总结规律，给出每台售货机饮料类的商品的标签，在此基础上进行标签拓展，依据标签生成完整的售货机图像，结合任务二中的热力图，可以更好的帮助经营者制定经营策略，掌握经营方向。

对于任务四，根据附件提供的数据，利用 python 对数据的处理和分析构建商业数据预测模型，对每台自动售货机的每个大类商品的 2018 年一月的交易额进行预测。

目　录

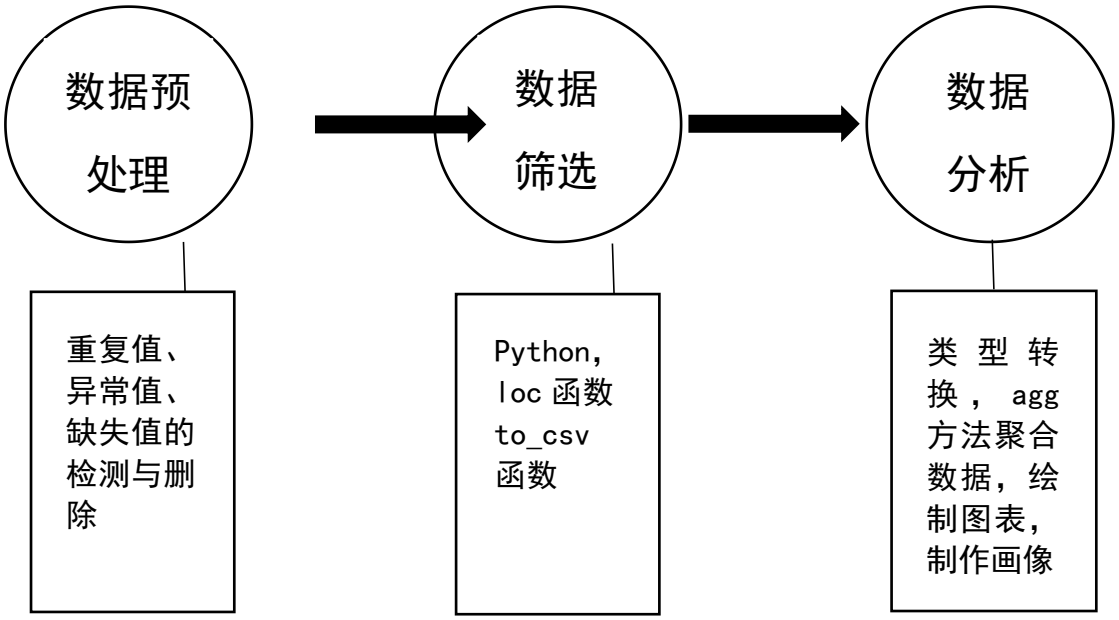
1、项目目标

本次数据挖掘目标是利用附件一给出的不同地点自动售货机 2017 年一整年的销售数据，利用 python 对其交易金额进行统计分析，达到以下两个目标：

（1）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，按其要求绘制对应图表，分析每台售货机 2018 年 1 月商品销量的预测问题。

（2）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。

2、分析方法与过程



图一：总流程图

本案例主要包括以下步骤：

步骤一：数据预处理。在题目给出的原始数据上进行一系列的清洗、去重等预处理方法，在此基础上进行数据提取和分析。

步骤二：数据筛选。对数据进行数据预处理后，把附件一的原始数据根据地点进行提取，这里采用了 loc 函数对数据进行筛选分类，在用 to_csv 函数对提取后的数据进行保存。

步骤三：数据分析。根据商品销售数据，对自动售货机进行商务数据分析，探讨销量、营业额、利润等之间的关系。绘制自动售货机的画像，分析售货机商品的热销程度，给出合理的营销意见。

步骤四：了解预测的含义，掌握预测方法，尝试预测 2018 年一月的销售数

据。

2.1 任务一的分析方法和过程

2.1.1 数据预处理

在 python 上利用 duplicates 函数、isnull 函数等对附件一的数据进行去重、去空等数据清洗操作，将原始数据中不合理的部分删除，再将剔除后的数据保存在原附件 1.csv 中。

2.1.2 数据提取

在任务一中要求计算每台售货机的交易情况，因此需要在完成预处理的步骤后，将附件 1 中的原始数据按地点进行提取并保存。步骤如下：

（1）根据附件 1 原始数据中“地点”一列的信息，使用 loc 函数对其进行分类筛选，得到 A、B、C、D、E 五个地点的自动售货机的数据。

（2）保存到 csv 文件中，文件名分别为“task1_1A.csv”，“task1_1B.csv”,…,“task1_1E.csv”。

2.1.3 数据统计

要对自动售货机进行商务数据分析，探讨销量、营业额、利润等之间的关系，就需要先进行对其销售数据的统计，步骤为：

（1）使用 sun 函数在 python 中对附件 1 的“实际金额”数据进行计算得出总的交易金额，再利用 shape 得知附件 1 中有多少行就可得知有多少订单量。结果如图所示。

	自动售货机
交易总额	286979.7
订单总量	70679

（2）将 A、B、C、D、E 五台售货机数据中“支付时间”的类型通过 datetime 函数从 object 改为 datetime64，再用 index 和 agg 函数按月份提取数据并计算其每月交易金额、每单平均交易额和日均订单量。结果如图所示。

A 售货机

	一月	二月	三月	四月	五月	六月
交易额	1509.7	440.5	914.3	1804.5	3385.1	6755.1
日均订单量	335	114	255	447	756	1669
每单平均交易额	4.506567	3.864035	914.3	4.036913	4.477646	4.047394
	七月	八月	九月	十月	十一月	十二月
交易额	1950.5	2236.9	4479.5	6292.4	5187	7587.1
日均订单量	476	666	1040	1565	1160	2003
每单平均交易额	4.097689	3.358709	4.307212	4.020703	4.471552	3.787868

B 售货机

	一月	二月	三月	四月	五月	六月
交易额	1373.6	602.3	957.9	2457.4	3681.2	7550.3
日均订单量	366	185	265	603	869	1856
每单平均交易额	3.753005	3.255676	3.614717	4.07529	4.236133	4.06805
	七月	八月	九月	十月	十一月	十二月
交易额	1518.6	3516.1	7207.3	8331.6	8669.9	8104.1
日均订单量	345	981	1745	2026	2031	2210
每单平均交易额	4.401739	3.5842	4.130258	4.11234	4.268784	3.667014

C 售货机

	一月	二月	三月	四月	五月	六月
交易额	1640.5	792	991.5	3232.3	3729.4	8472.2
日均订单量	379	207	263	734	789	1882
每单平均交易额	4.328496	3.826087	3.769962	4.403678	4.726743	4.5017
	七月	八月	九月	十月	十一月	十二月
交易额	3047.1	4927.2	7429	9469.7	8456.7	9380.5
日均订单量	764	1259	1678	2216	1943	2379
每单平均交易额	3.988351	3.913582	4.427294	4.27333	4.352393	3.943043

D 售货机

	一月	二月	三月	四月	五月	六月
交易额	956.4	435.5	826.7	1679.1	2392.1	4187
日均订单量	259	141	192	443	564	1040
每单平均交易额	3.692664	3.088652	4.305729	3.790293	4.241312	4.025962
	七月	八月	九月	十月	十一月	十二月
交易额	1340.8	2371.3	3833.1	4606.7	4673.4	5941.2
日均订单量	317	715	983	1186	1210	1663
每单平均交易额	4.229653	3.316503	3.89939	3.884233	3.862314	3.57258

E 售货机

	一月	二月	三月	四月	五月	六月
交易额	1656.8	938.7	1507	3723.1	5699	9899.7
日均订单量	354	258	350	895	1292	2593
每单平均交易额	4.680226	3.638372	4.305714	4.159888	4.410991	3.817856
	七月	八月	九月	十月	十一月	十二月
交易额	3186.4	6722.5	17054.3	10208.6	21501.8	13557.5
日均订单量	813	1767	4134	2777	5020	3252
每单平均交易额	3.919311	3.804471	4.125375	3.676125	4.283227	4.168973

2.2 任务二的分析方法和过程

2.2.1 数据筛选和统计

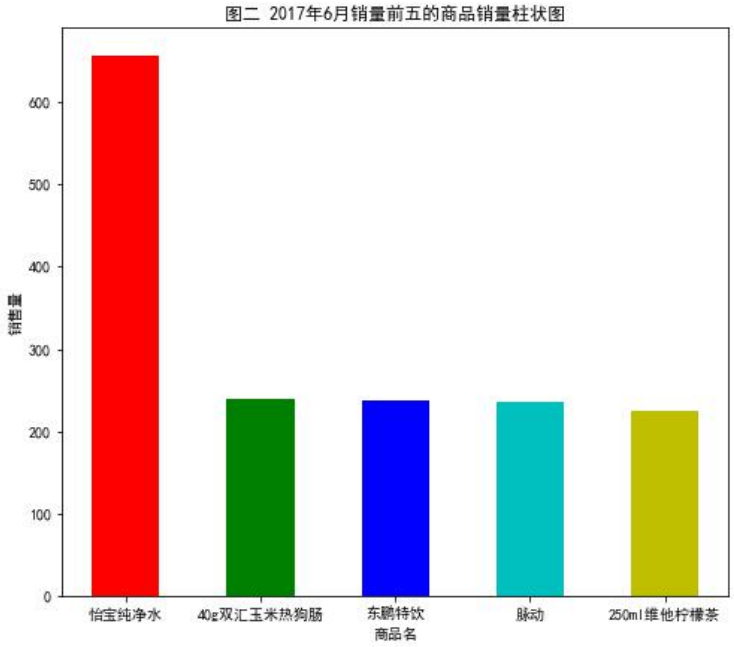
（1）在附件 1 里的 2017 年 6 月份的销售数据中，对每种商品出现的次数进行计数，通过排序得出排名前五的商品，再将其可视化成直方图的形式，更加直观且清洗的了解到销量前五的商品销售情况。因为 value_counts（）拥有自动从高到低的排序功能，所以在 python 使用 value_counts（）的方法，可以直接得出排名前五的商品名称及其销量，缩减了运算量和运算时间。

（2）对各个自动售货机进行分月计数，得出每台自动售机每月的总交易额，再计算交易额的月环比增长率，可知在哪一个月份时，自动售货机的使用情况最好，即知道什么时候是淡季和旺季，使经营者更好的安排自动售货机的供给频率。由于在任务一中已经计算出了五个自动售货机的每月总交易额，将得出的这些数据导入 Excel 表格中，再使用 Excel 自带的函数计算功能，轻而易举的得知五台售货机交易额的月环比增长率。

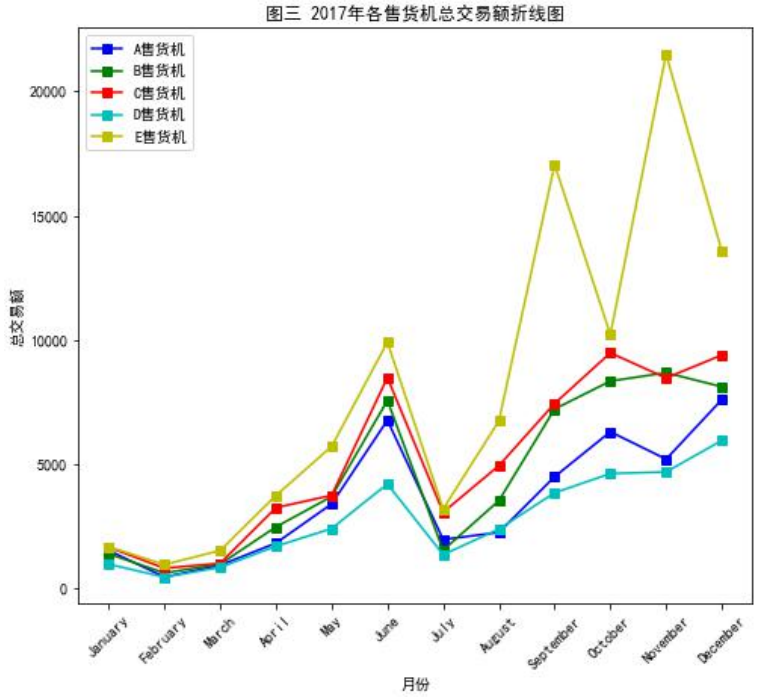
（3）计算每台自动售货机毛利率占总毛利润比例。首先在 python 中读取附件 2 的数据，再将附件 2 和附件 1 合并在一起，然后分别筛选出非饮料类和饮料类的实际金额和地点，再使用 for 循环，将不同的地点的非饮料类、饮料类的总金额计算出来，分别乘以毛利率，得出每台自动售货机的毛利率，最后在 plt.pie 函数中输入五台自动售货机的毛利率，就可得出五台自动售货机毛利率占总毛利润比例的饼图。

2.2.2 数据可视化及分析

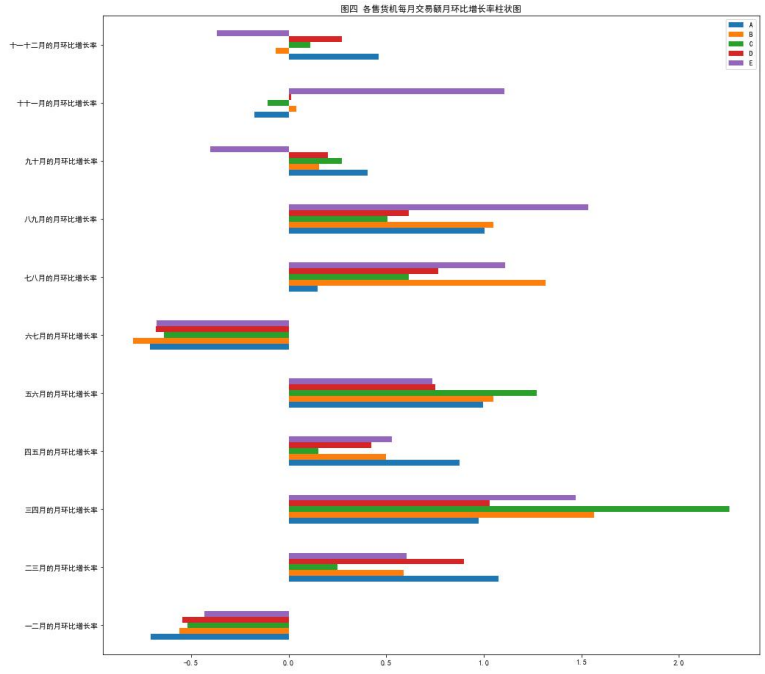
（1）从图二中可以看出，在 6 月这一个月中纯净水的销量最高，其余前四的商品销量相差不大。



（2）在图三中可以看出，一到七月份的五台自动售货机的月总交易额折线的波动趋势相似，营业额也相差不大，但八月过后，E 地售货机的交易额增长得非常快，特别是九月和十一月，营业额达一万五以上了，而其它售货机则都在一万以下。所以经营者应特别注意 E 地的售货机，在九月份和十一月份时要提高商品的供给频率和供给量。要根据各地自动售货机每月的经营额和波动合理安排各地自动售货机的供给时间段，商品供给频率和数量，减少不必要的供给时间。

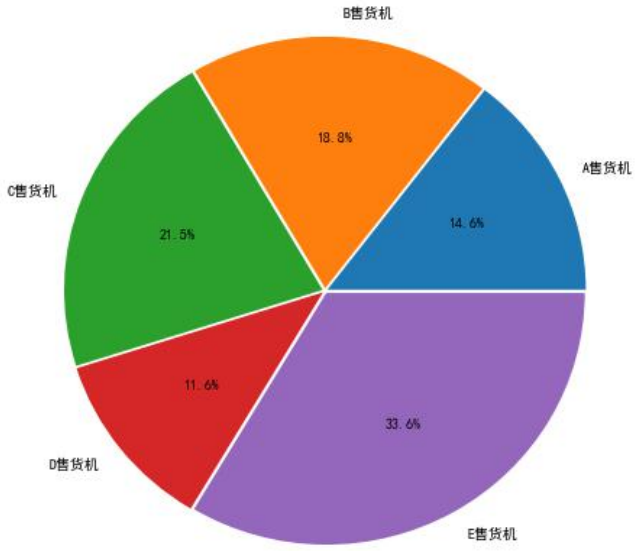


（3）而在图四的月环比增长率中，我们可以看到一二月和六七月的月环比增长率是负值，这说明二月份和七月份的销售情况都不怎么好，才会出现负增长，所以在这两个月份的商品供给可以适当地减少。而 B 地和 C 地的三四月的月环比增长率、E 地的八九月的月环比增长率相对而言较高，所以在月份交换之间，应及时注意补充自动售货机的库存量，以避免影响该时间段的销售情况。



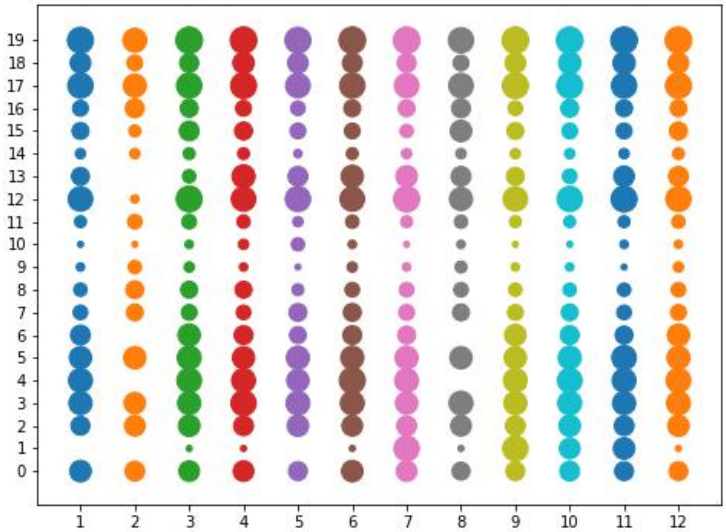
（4）通过图五的毛利润占比饼图可以看出，五个地点中 E 地的自动售货机所占的毛利润最高，在总体的毛利润中占百分之三十三点六，B 地和 C 地的售货机毛利润所占比在其中属于中等水平，而 A 地和 D 地的毛利润所占比是五个地点最少的。结合图二折线图和图三柱状图的波动趋势，在五个地点的售货机中，经营者的资源要根据以上情况，合理的倾向于 E 地，对 E 地自动售货机的供给频率、供给量和商品种类要比其余四个地点适当提高，跟上 E 地的销售情况，才可以获得更高的交易额，避免出现供给跟不上需求，造成营业流失的情况。

图五 2017年每台售货机毛利润占总毛利润比例的饼图



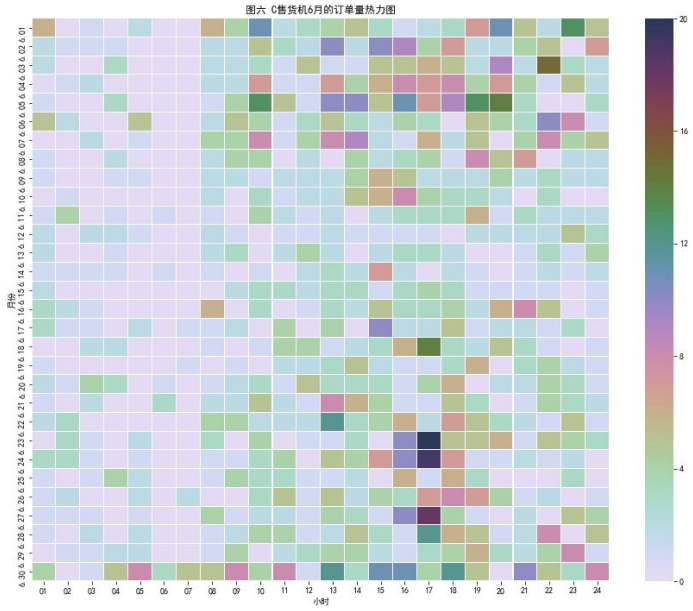
（5）先以商品为中心合并附件 1 和附件 2，在利用 for 循环得出每个二级类目商品的每个月交易额均值，气泡大小与每个月各二级类目商品的交易额均值大小有关。绘制气泡图如下：

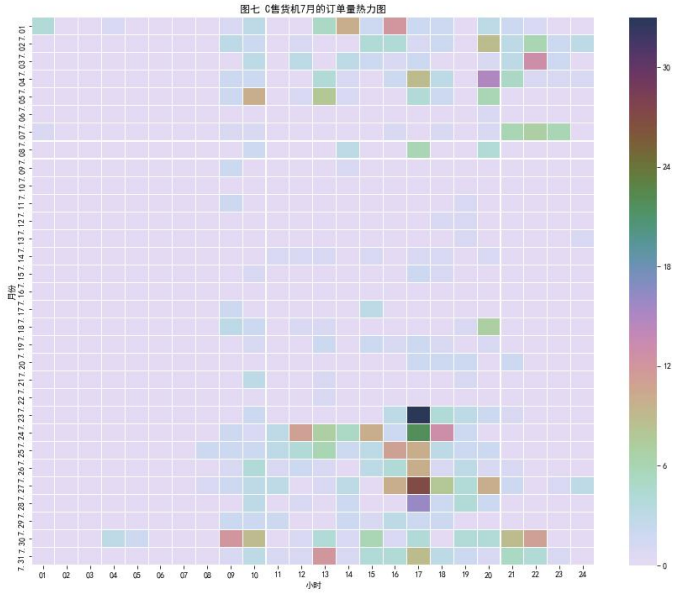
（横轴表示月份，纵轴表示商品，气泡越大代表该商品在本月交易额均值越高。）

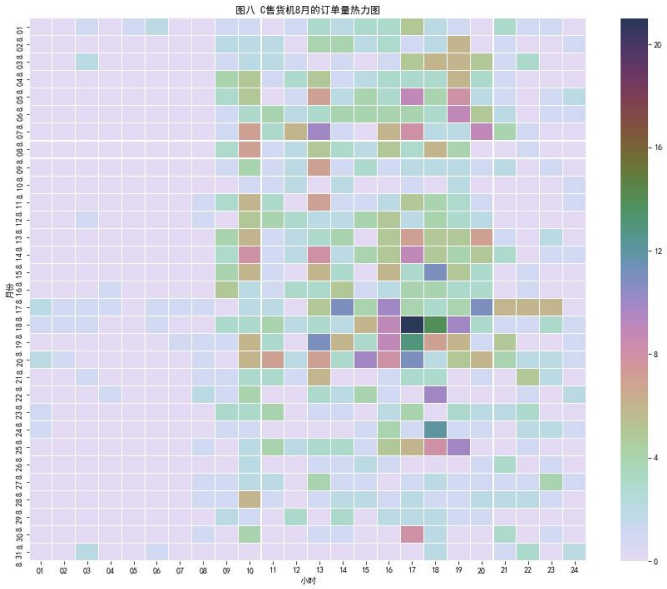


（6）以下的图六、图七和图八的热力图将 C 地自动售货机六月、七月、八月三个月的订单量数据分布通过不同颜色区块可显性的、直观的呈现出来，给经营者优化和调整自动售货机提供了有力的参考依据。同时，热力图还能告诉我们，自动售货机的哪个日期和哪个时间段的顾客量大，能帮助经营者分析安排供货时间和售货机检查时间。

在这三个月的数据中，我们发现 17 点这个时间段的订单量是较多的，C 地七月中旬售货机的销量低迷，产品挤压情况较严重。







2.3 任务三的分析方法和过程

（1）分析自动售货机饮料类和非饮料类的商品销售情况，统计出饮料类和非饮料类每种商品（包括同种商品不同口味）的销售数量，并为其贴上标签。标签可帮助经营者选择自动售货机投放商品的种类问题，而商品的销售数量则能帮助经营者合理的制定出每种商品的供给方案。（其中标签分别为畅销、正常、滞销这三个。）

（2）标签的区分方法：按照百分比计算，将所有饮料类商品的总销售数量看成一个百分比，将其除以 3，得出两个临界百分比，分别是 33.33%和 66.66%，将这两个百分比，饮料类和非饮料类的商品销售数据代入到 python 的 numpy.percentile 函数中就可计算出这两个临界百分比的数值，按照其数值，使用 for 循环就可以将饮料类和非饮料类的商品分为三个层次。

从以上方法分析出的结果看，饮料类商品的销售情况要比非饮料类的好，在饮料类畅销商品中有 14 样商品的销售数量破千，5 样商品破两千，1 样商品破四千，而在非饮料类中，只有 2 样商品破千，由此看来，自动售货机内饮料类商品比非饮料类要受欢迎，根据此情况，经营者可下架一部分饮料类和非饮料类的滞销商品，并适当的增加饮料类畅销商品的数量，使售货机维持并尽可能的提高其销售金额。

2.4 任务四的分析方法和过程

2.4.1 预测原理

（1）连续性原理：事物的发展是按照一定规律进行的，而这种规律一般都是连续存在的。

（2）类比性原理：类似事物的发展趋势也类似，可以通过类似事物的发展来预

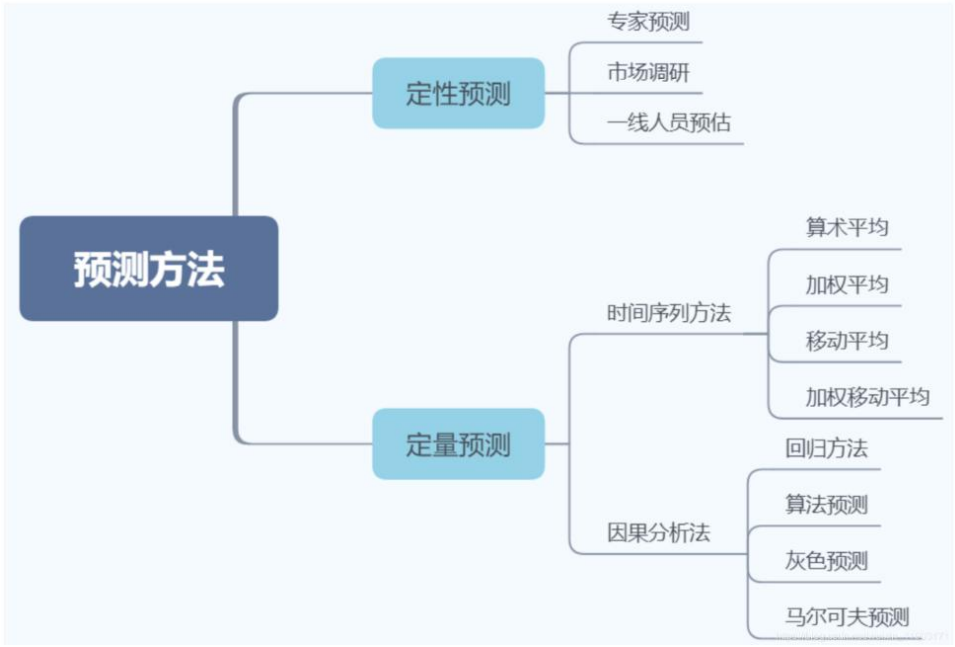
测。

（3）相关性原理：一个结果指标由若干因素影响，通过分析影响因素来预测结果。

（4）概率推断原理:根据以往表现求出随机事件出现各种状态的概率，然后进行预测。

2.4.2 预测要求

（1）常用预测方法

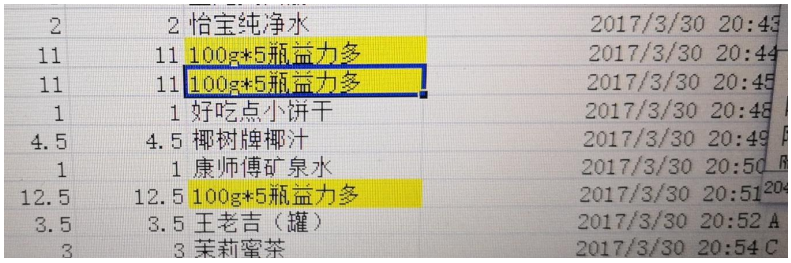


（2）预测问题

用以上预测方法建立预测模型都是可行的，但是本案例却无法使用。原因如下：

原因一：数据不完整。附件 1 中三月份数据严重缺少，仅仅只有几天的数据，而不是整个月份的数据，如果用这份数据去预测 2018 年 1 月的未来销量，得出的结果也是不正确。

原因二：数据不严谨。我发现在附件 1 的“实际金额”这一栏的数据飘忽不定，同一种商品在同样的售货机上却又不同的价格，这样会导致数据预测结果不稳定。



原因三：数据量少。本案例中只给 2017 年 12 个月的数据，并没有给出 17 年之前的数据，数据量过少，会导致数据预测结果不准确。

因此，经营者应给出至少 5 个年份的数据，并数据要严谨，不能缺少，才能更好的使用预测模型，得出可信度较高的预测结果。