**超市数据分析报告**

# 一、数据基本情况及预处理

## 1.1数据基本情况

报告中对数据的处理，均使用jupyter notebook。

首先，将CSV文件导入，数据表“item”中共有1,732,761条数据，10个数据项。

统计不重复商品编码，得知该超市共有19,271种商品；其中出现次数即被购买次数最多的前十名商品及次数如下表1：

表1 出现最多的前十名商品编号及次数

|  |  |
| --- | --- |
| 商品编码 | 出现次数 |
| 205540 | 12622 |
| 21835 | 10459 |
| 194690 | 9548 |
| 72811 | 7747 |
| 178462 | 7373 |
| 197532 | 7289 |
| 205437 | 6725 |
| 1557 | 6041 |
| 109584 | 5932 |
| 216571 | 5675 |

根据不重复的店铺编码，此次统计中该超市共有24家门店；

表2 门店编码及出现的次数

|  |  |
| --- | --- |
| 门店编码 | 出现次数 |
| A001 | 193822 |
| A004 | 46005 |
| A006 | 77906 |
| A007 | 49326 |
| A008 | 106171 |
| A009 | 56326 |
| A010 | 49291 |
| A012 | 63749 |
| A013 | 81522 |
| A016 | 56266 |
| A017 | 28789 |
| A018 | 46569 |
| A019 | 51289 |
| A021 | 62889 |
| A022 | 44130 |
| D002 | 242492 |
| D003 | 22249 |
| D004 | 76548 |
| D005 | 50408 |
| D006 | 70071 |
| D007 | 117804 |
| D008 | 59282 |
| D009 | 45032 |
| D010 | 34825 |

不重复的购物篮编码，共有281,211个。正常情况下，一个购物篮编码对应一次购买行为，但是通过浏览数据，同样的购物篮编码对应一次或几次购买行为，如编码为62087的购物篮，在数据中共出现82次，这82次出现在不同的时间甚至门店。因此，购物篮编码不能单独用作统计交易次数的依据。

表3 购物篮编码情况

|  |  |
| --- | --- |
| 购物篮编码 | 出现次数 |
| 62087 | 82 |
| 37508 | 68 |
| 147354 | 68 |
| 208347 | 66 |
| 36227 | 66 |
| ... | ... |
| 170019 | 1 |
| 322103 | 1 |
| 165921 | 1 |
| 241844 | 1 |
| 6147 | 1 |

在数据项中，有一个数据项是打折类型，通过统计，共有9种打折类型，分别是n,p,m,v,x,q,Z,s,h。

表4 打折类型

|  |  |
| --- | --- |
| 打折类型 | 出现次数 |
| n | 1206812 |
| p | 354010 |
| m | 153875 |
| v | 11851 |
| x | 4530 |
| q | 1458 |
| Z | 202 |
| s | 22 |
| h | 1 |

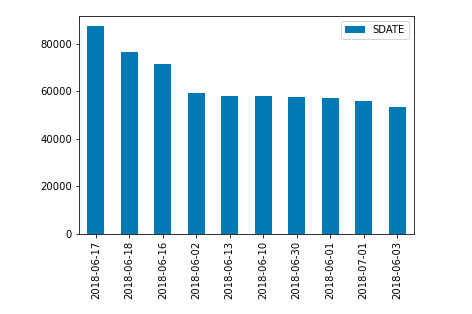
通过CSV文件名称或者数据统计可知，该数据集是超市从2018-6-1到2018-7-5，共35天的交易数据，其中发生交易次数最多的前十天情况，如下图1所示：

图1 交易次数最多的十天

## 1.2数据预处理

通过大致浏览数据，发现数据中存在一些异常数据，需要对数据进行整理清洗。

首先，筛选出数据中有空值的行，经过筛选，数据中不存在空值。

在"ITEM\_QUANTITY"、"ITEM\_SPEND"和"DISCOUNT\_AMOUNT"三列中，数据符号有正有负，根据实际情况及相关判断，将这三列数据全部通过取绝对值变为正数，方便后续分析。

通过公式计算实际折扣，将计算所得作为新的列添加到原数据中。为了比较实际折扣与表中已知的折扣，将两列数据均保留两位小数。筛选出实际折扣与已知折扣不相等的数据行，共5971行，异常数据占总体数据不足1%，删除后样本量仍然足够，对后续数据分析影响不大，故删除这些异常值。

经过预处理，现样本数据为1,726,790条。