零售超市商品数据分析

第一周主要进行了数据集的导入、数据查看、缺失数据的查看、异常值的查看与清洗、整体销售情况分析、商品维度分析和店铺维度的初步分析。

# 一、数据集描述

本文数据集为真实零售超市的真实数据，该数据集包含2018年6月1日-2018年7月5日的公司网上零售的交易信息。

表 1 basket数据表字段与描述

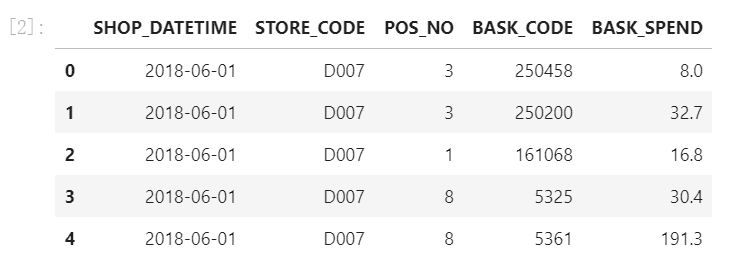
|  |  |
| --- | --- |
| **字段** | **描述** |
| SHOP\_DATETIME | 订单日期 |
| STORE\_CODE | 商店编号 |
| POS\_NO | POS机编号 |
| BASK\_CODE | 订单编号 |
| BASK\_SPEND | 订单金额 |

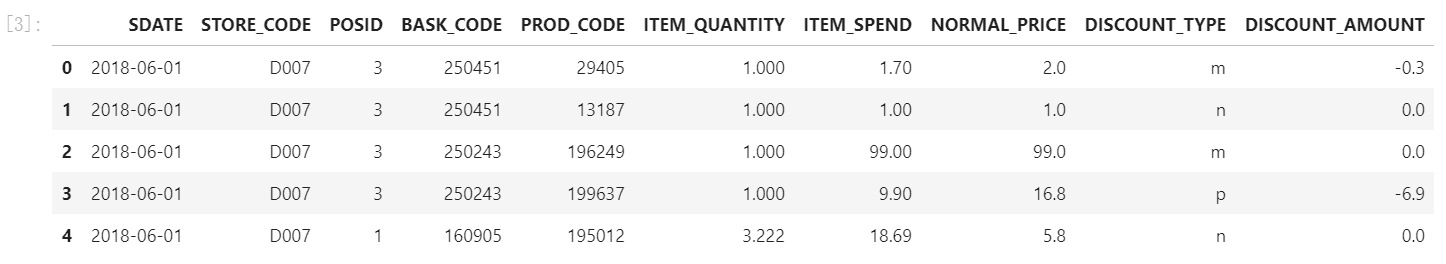
表 2 item数据表字段与描述

|  |  |
| --- | --- |
| **字段** | **描述** |
| SDATE | 订单日期 |
| STORE\_CODE | 商店编号 |
| POSID | POS机编号 |
| BASK\_CODE | 订单编号 |
| PROD\_CODE | 商品编号 |
| ITEM\_QUANTITY | 商品数量 |
| ITEM\_SPEND | 商品实际价格 |
| NORMAL\_PRICE | 商品单价 |
| DISCOUNT\_TYPE | 折扣类型 |
| DISCOUNT\_AMOUNT | 折扣金额 |

# 二、导入数据

导入basket数据和item数据并简单查看数据

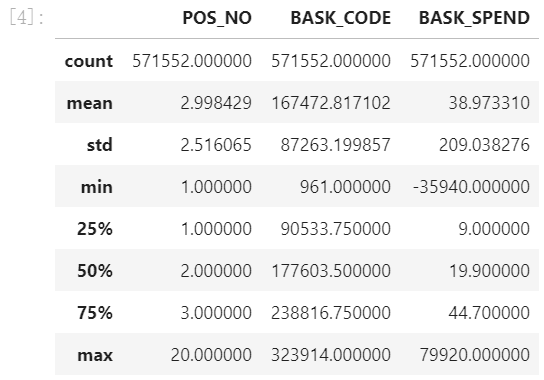
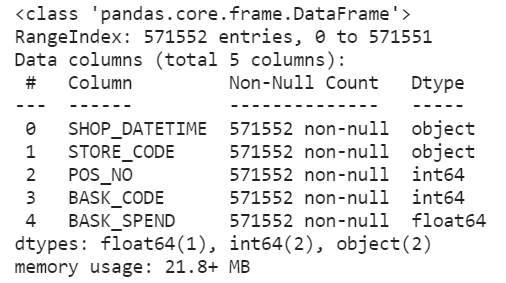




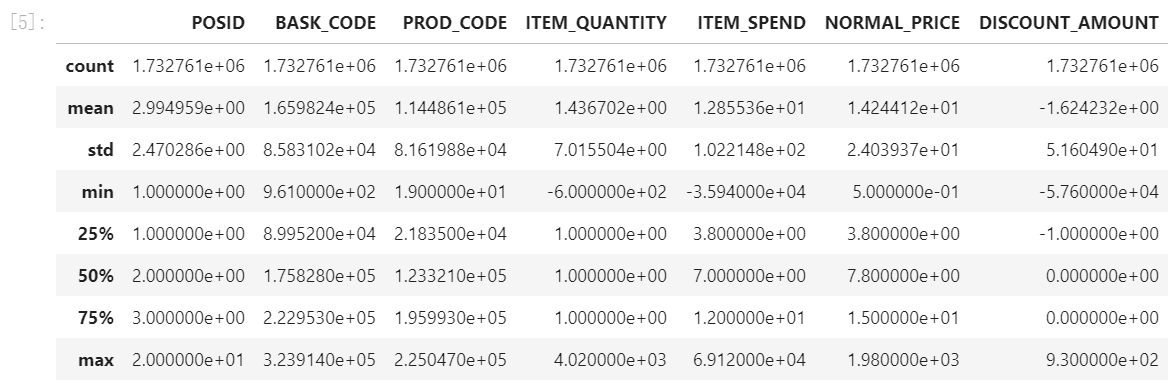
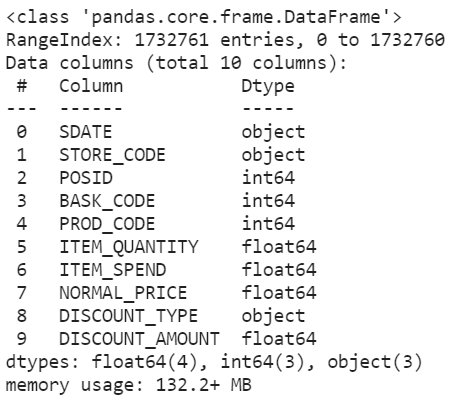
# 三、数据查看与清洗

## **1、查看数据特征**

## **（1）查看basket数据总体特征**



## **（2）查看item数据总体特征**



## **（3）小结**

从图片中初步可以看出数据集可能存在的异常数据以及需要进行初步处理的地方有以下几点：

①两个数据表中的SDATE字段均需要做数据类型转换；

②Basket表中的BASK\_SPEND字段有负值，即订单金额为负；

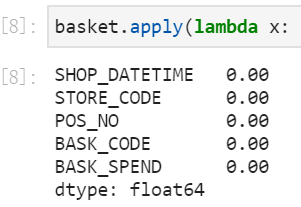
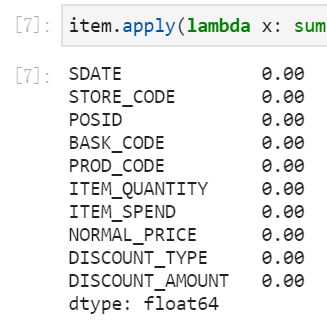
③item表中的ITEM\_SPEND字段有负值，即商品实付金额为负；

④item表中的ITEM\_QUANTITY字段有负值，即商品数量为负；

⑤item表中的DISCOUNT\_AMOUTN字段数据有正值，即折扣后金额比折扣前金额高。

## **2、缺失数据**

## **（1）统计缺失率**

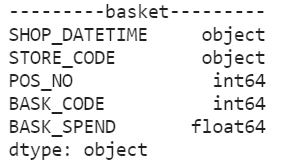
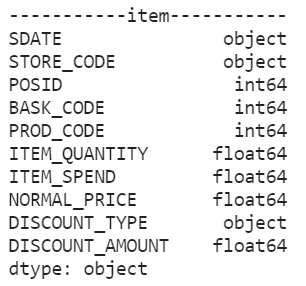


## **（2）缺失数据分析**

数据不存在确实情况，不需要对缺失数据进行填充。

## **3、数据类型**

## **（1）查看数据表数据类型**



## **（2）转换数据类型**

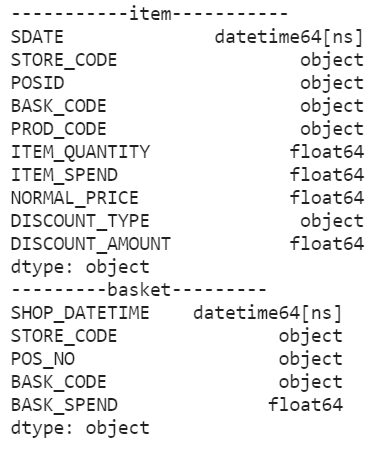
数据类型中的object表示如果一列中含有多个类型,则该列的类型会是object,同样字符串类型的列也会被当成object类型，因此数据类型中object类型需要调整。此外为方便计算，涉及编号的数据类型应为String类型，SDATE数据类型应为日期类型，数值类型由于涉及零售的金钱问题，也应都调整为两位小数。

①调整编号数据类型

②调整日期数据类型

③将金钱调整为两位小数

④调整后数据类型



## **4、查看异常值并删除**

根据上一小节对数据的初步查看发现的问题进行操作，在上一小节中发现数据表中商品数量、订单总金额、商品实付金额均有负值，且折扣金额为正，折扣金额的计算可能受到金额为负的影响，此外，数据也可能存在空值。因此，异常值的删除主要有以下几个操作：

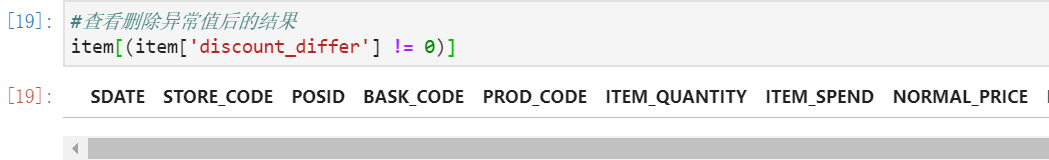
①将商品数量、订单总金额、商品实付金额为负值的调整为正值；

②新建销售总金额字段：销售总金额=单价\*数量

根据修改后的数据进行折扣金额的计算：折扣金额=销售总金额-实际付款金额（ITEM\_SPEND），与实际折扣金额不同的可以判定为异常数据，进行删除操作；

③删除过后分析是否还存在折扣金额为正的数据，若存在，也进行删除操作。

## **5、删除过后对数据进行查看**



最后发现不存在折扣金额与实际折扣不同的数据，可以判定无异常值。

# 四、数据分析

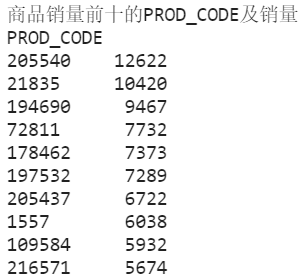
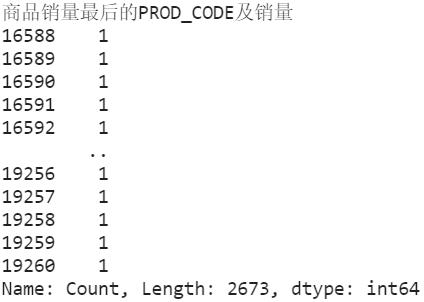
## **1、总体销售情况分析**

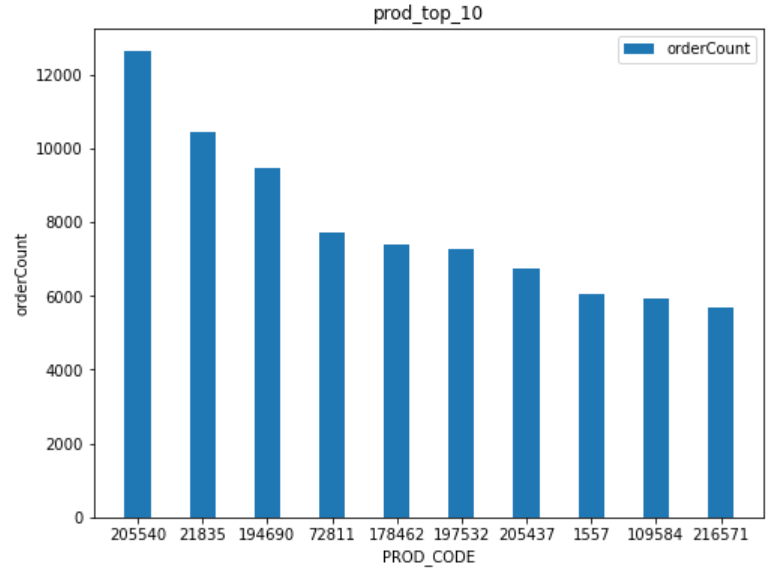


由总体销售情况分析可知，商店在2018年6月1日-2018年7月5日总订单数有280878单，总销售额为22216624.17元，有流水的商品数共19261个。

## **2、商品维度分析**

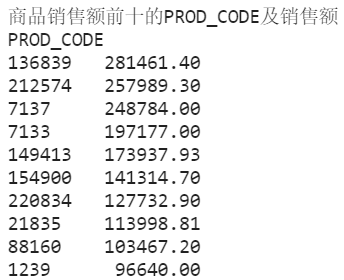
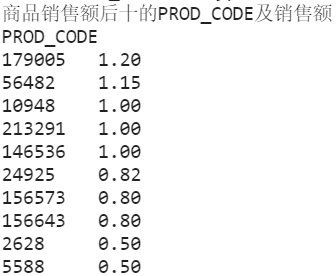
## **（1）商品销量分析**

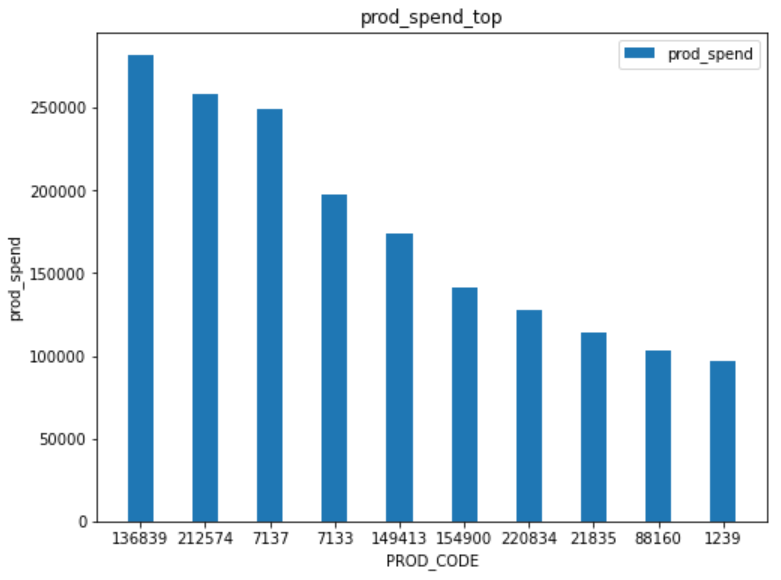
 



本小节分析了商品销售量前10和商品销售量最后的商品数量，此外将商品销售量前10的商品绘制柱状图进行分析。由图表可知，前3销售量的商品很明显比后面商品销量多，差别至少为2000以上，而商品销量最后的商品在计算过程中发现有很多商品销量为1，将所有销量为1的商品提取出来，共有2673个，这2673个商品在这一段时间只销售了1个商品，可能为商品为不常用商品，也可能是商品本身有其他问题。

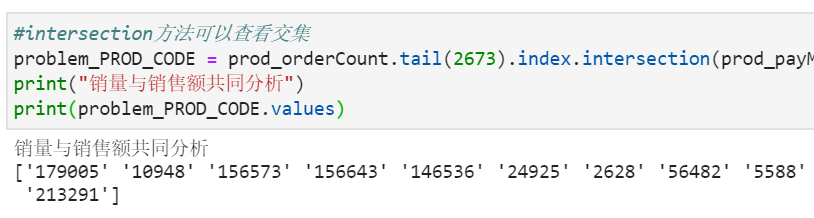
## **（2）商品销售额分析**



本小节主要分析了商品销售额前10和商品销售量后10的商品，此外将商品销售量前10的商品绘制柱状图进行分析。由图表可知，前3销售量的商品很明显比后面商品销售额多，差别至少为50000以上，而商品销量最后的商品仅仅在1元左右，这些商品无论是数量还是销售额都很少，商品可能存在问题。

## **（3）查看销售量与销售额关系**

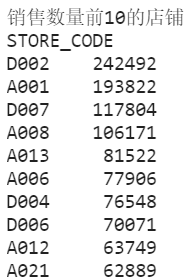


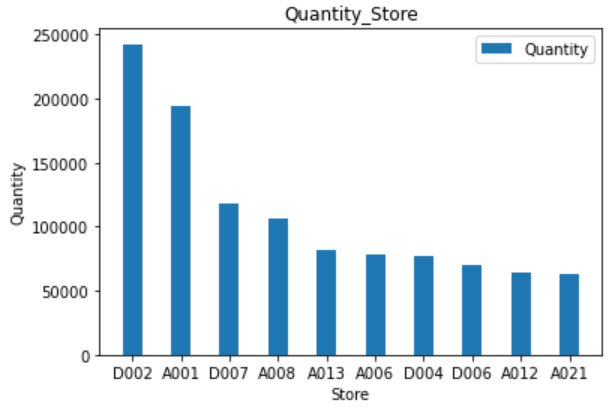
由于商品销售量最后的销售量为1，数量共2673个，因此查看销售量最后2673个商品中同时销售额在后10的商品，发现销售额最后的商品均为销量为1的商品，查看了商品销量和销售额之后，就可以查看销量和销售额都排名靠后的商品是哪些，找出之后再看具体是哪些商品，找出原因，考虑如何优化或者是否要下架。

## **3、店铺维度分析**

进行店铺维度的分析可以分析各店铺销售情况，判断哪些店铺销售情况不好，考虑是否需要对店铺员工进行培训或裁员。

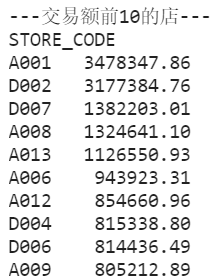
## **（1）店铺销售量情况分析**





本小节分析了商品销售量前10的店铺，可以发现商品销售数量前2的店铺销售量远超过后面的店铺，超过150000个商品，说明D002和A001两个店铺商品数量销售情况很好。

## **（2）店铺销售额情况分析**





本小节分析了商品销售额前10的店铺，进行图表分析，可以发现A001和D002两个店铺销售额远超其他店铺，至少超过了2000000元，根据上小节分析，这两个店铺销售量与销售额均远超其他店铺，说明两个店铺销售情况良好。