项目详细方案

一、题目分析

近年来，互联网技术发展迅速，电子商务行业也踏上了发展的快车道。在各平台规模不断扩大、商品数不断增加的同时，一些不正当的经营行为，例如虚标价格、刷单行为也随之出现，严重违反了电商法，因此我们需要对这类商品数据进行准确识别。

各类网上零售平台提供的商品，我们可以利用其提供的商品价格、销量、类目参数、店铺等信息，对这些商品标注其中价格异常和销量异常的商品。根据题目要求并结合实际情况我们考虑价格和销量高于正常值的商品数据为异常数据。又因为不同商品的价格和销量并不具备统一的规律，因此我们分别考虑价格异常和销量异常。

对于价格异常，我们主要通过价格指标和其余参数指标建立相应模型。初步分析发现同一类商品价格相差不会很大，因此我们计算出价格的平均范围并由此划定出合理价格区间，然后再通过概率分布来筛选出异常数据。

对于销量异常，我们主要通过销量指标来分析异常数据。考虑到商家刷单只会在某段时间内进行，所以我们通过分析该店铺几个月的数据得出异常数据。

为了得到更多的信息和更有针对性的特征，我们接下来对数据进行了深度的探索。

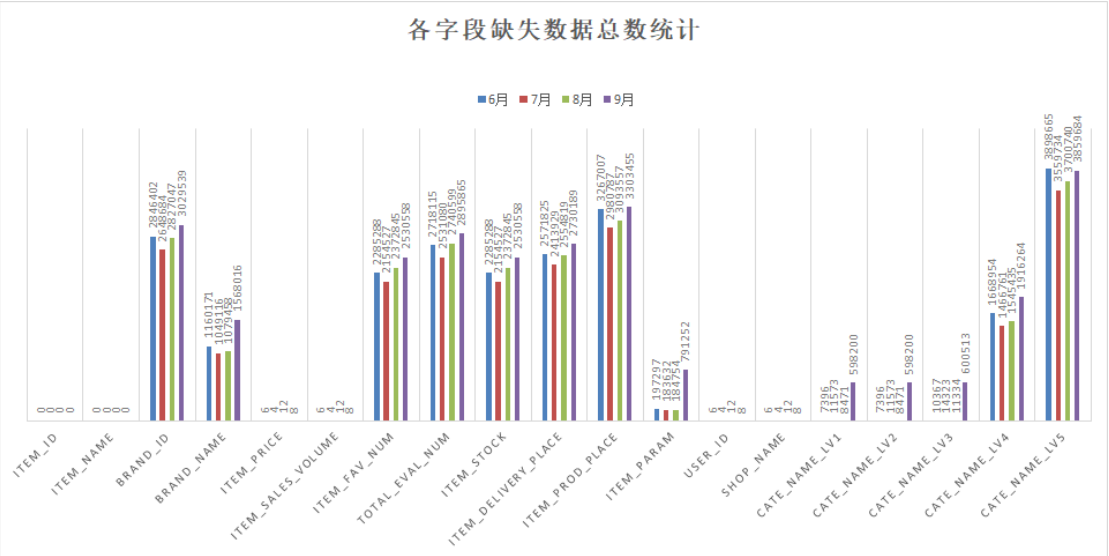
二、数据理解

2.1 探索新方案

由于数据量较大，导致我们无法使用传统数据分析工具对其进行处理，并且考虑到可扩展性以及灵活性问题，故我们在技术选型的时候采用的是基于内存的大数据分布式框架 spark 进行处理，以便我们可以更加高效的处理分析数据。

2.2 原始数据分析

首先我们对商品数据进行分析，统计了各字段缺失数据的总数：



我们发现 BRAND_ID、ITEM_SALES_VOLUME 等字段缺失数量超过近一半， 并且绝大部分字段我们无法对其进行填充，例如 BRAND_ID（商品 ID）、ITEM_FAV_NUM（库存量）等。于是我们尝试从其余缺失较少的字段中提取相关特征 ITEM_ID、ITEM_NAME、

ITEM_PRICE 、 ITEM_SALES_VOLUME 、 USER_ID 、 SHOP_NAME 、 CATE_NAME_LV1 、 CATE_NAME_LV2、CATE_NAME_LV3 等，但是我们发现这些字段蕴含的信息量较少，无法很好的完成异常检测任务。

2.3 类目划分

我们对其余字段进行观察分析，我们发现类目字段中前三级类目字段缺失数量较少，后两级类目字段缺失数量较多，但是经过查看具体数据后发现类目字段的缺失并不是真正的缺失。

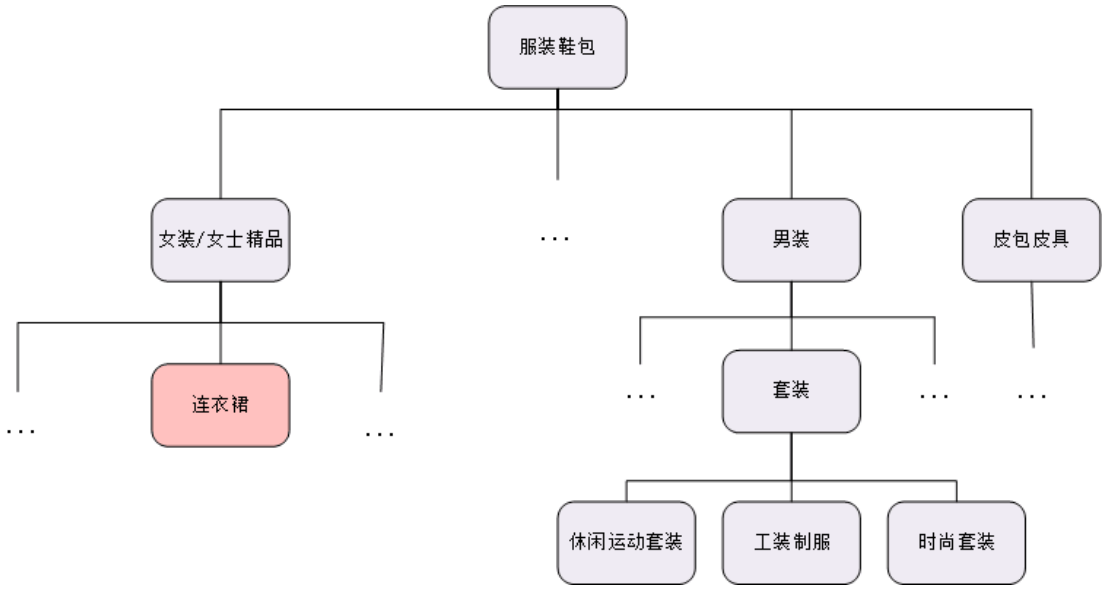


图 2.1 商品类目结构图

从以上结构图可以看出在电商平台中，每件商品会被划分到不同的类目，平台首先会将一件商品划分到一个大类中，然后在一个大类里一层层下分到某个小类，直到不可细分为止。假如这件商品只被划分了三次就不可细分，那么其后的两级类目就为缺失。

故我们最后选取的字段有 ITEM_ID、ITEM_NAME、ITEM_PRICE、ITEM_SALES_VOLUME、USER_ID 、 SHOP_NAME 、 CATE_NAME_LV1 、 CATE_NAME_LV2 、 CATE_NAME_LV3 、 CATE_NAME_LV4、CATE_NAME_LV5、ITEM_PARAM。

但是在部分字段中仍有数据缺失，由于这部分数据缺失数量占比十分小，故可以直接舍去。

三、价格异常：

3.1 思路：

经查阅资料发现商品价格受到多方面的影响：价值决定价格、供求关系影响价格、国家政策影响价格、消费心理影响商品价格、地域条件、生产条件等。

价值决定价格。价值是价格的决定性因素，商品的价值越大，商品的价格就越高，不同类别的商品由于价值差异无法直接进行比较。

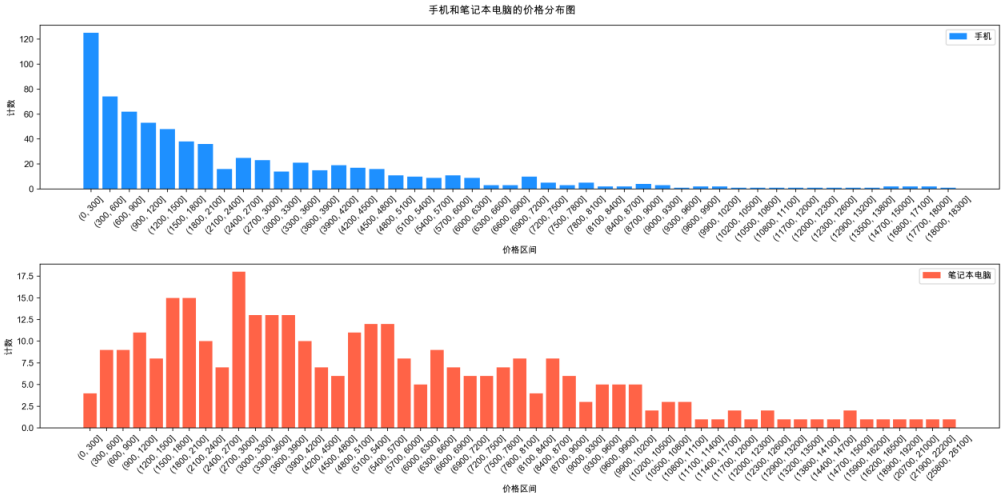


图 3.1 价格分布趋势图

如上图，可以看出对于不同类别的商品会有不同的价格区间。并且同类商品的价格会有一个大致的分布区间，于是我们在判断一件商品的价格是否存在异常的时候，我们可以先确定该件商品的实际价值，然后找到该件商品的价格的范围，若某件商品的定价超出其实际价格浮动范围，则判定这件商品属于价格异常。经过观察和分析发现如果我们先对商品按照 5 个类目进行分类，我们就可以得到同类商品，对于同类商品，我们就可以使用上述方法判定价格异常。

3.2 具体方法

3.2.1 选择变异系数评估波动大小

我们用 5 个类目字段对每一件商品分好类之后，再通过分析每一个类别的价格波动，发现有一些类别的价格波动范围不是特别大，

$$CV = \frac{\sigma}{\mu}$$

其中CV表示变异系数，σ,μ分别表示该类商品的方差和平均值。我们选取统计学中的变异系数来衡量波动大小，通过处理得出每个类别的变异系数并确定一个变异系数阈值来筛选出可能存在价格异常的数据。

3.2.2 用对数正态分布拟合确定阈值

在上述工作后，我们可以得到存在异常数据概率较大的类别。之后我们再分析这些类别中商品价格的分布，发现同一类别的商品的价格相比于其他分布而言更符合对数正态分布

	sumsquare_error	aic	bic	kl_div
lognorm	0.000008	1971.939735	-449180.1037	inf
chi2	0.000022	2064.05363	-428673.5455	inf
t	0.000028	2117.57193	-423317.0633	inf
exponpow	0.000031	2038.896663	-421474.7337	inf
laplace	0.000048	2498.870315	-412454.0486	inf

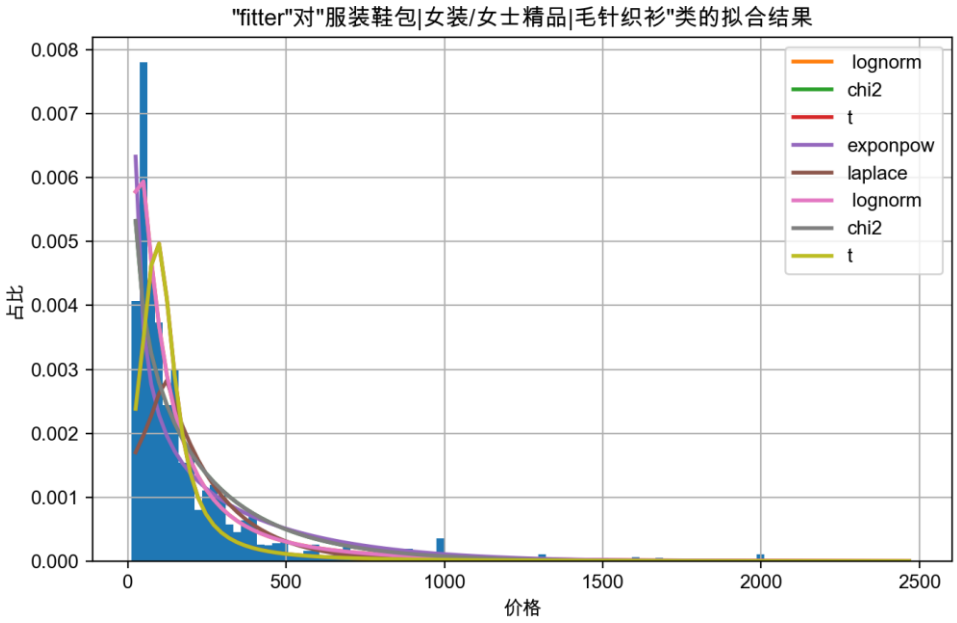


图 3.2 服装类价格拟合图

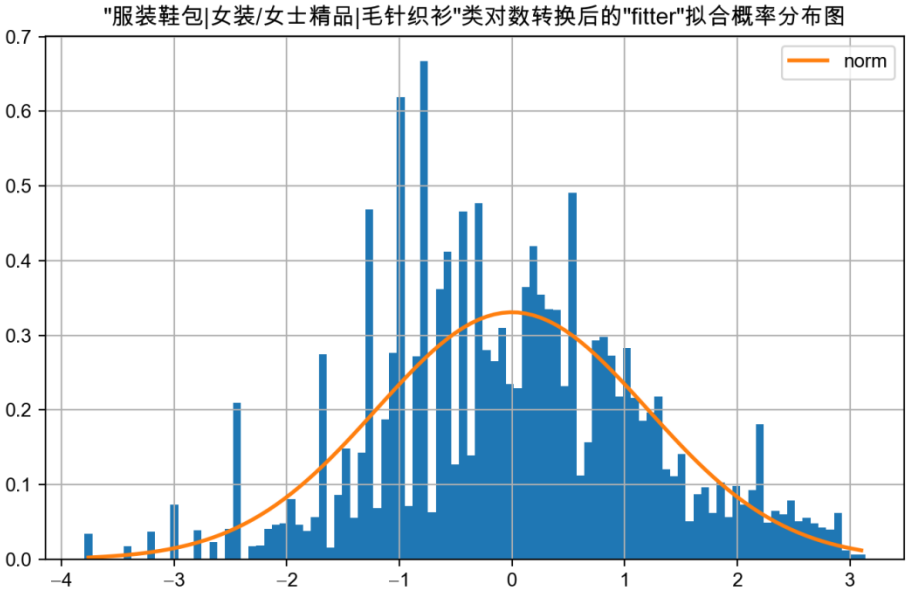


图 3.3 服装类价格概率分布图

分析上图结果我们可以看出基本符合对数正态分布。通过查阅资料，我们发现对数正态分布可以转换成正态分布，再基于统计学的规律，超过 3e 之外的数据为异常数据，因此我们拟合出某类商品的对数正态分布后再转换成正态分布，再基于统计学 3e 原则，得到右侧阈值，我们定义高于这个阈值的价格为异常价格。

3.2.3 用价格聚类来确定一个评分

（1） 聚类算法

我们对同一类商品的价格进行聚类，求出每个点到离他最近的 k 个点距离和的平

均（简称平均距离），以此为每一件商品的评分。

$$\sum_{item_i \in A} \frac{1}{k} \sum_{item_j \in B_i} | abs(\text{value}\_item_j - \text{value}\_item_i) |$$

其中 A 表示物品集合，B_i 表示离物品 i 最近地 k 个物品，$value\_item_i$ 表示物品 i 的价格。

然后我们取这类别中商品个数的 10%为 k 值，这种方法的复杂度为(n^2)/10。考虑到数据量过大，通过这种方法可能会花费较多的时间，因此我们对这个算法进行优化。

（2） 优化算法

通过“桶+二分+前缀和”的方法将复杂度优化到 O(m*logm) （m 表示这一类商品中价格的种数，且 m<n)。

首先用桶来存储每种价格出现多少次，对于每种价格，我们们二分边界价格距离，用前缀和的思想 O(1)验证区间的商品个数是否大于等于 k，最后再进行边界判断，由此种价格的复杂度为 log(m)，总的复杂度为 mlog(m)，所以可以较快的处理大规模的数据。

四、销量异常：

通过观察分析大量的数据，我们发现部分商家的销售数据会在某个月内迅猛增长，且该月的销售量远超其他月份，并且在销量猛增后迅速下降，如下图所示：

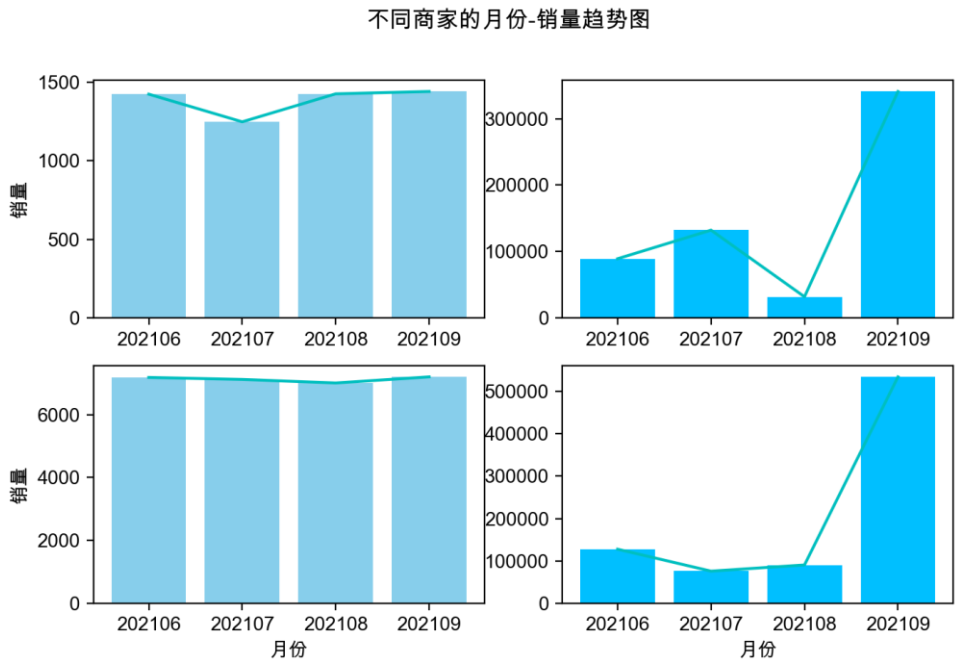


图 4.1 商家各月份销量趋势图

从上图可以看出右边两个商家的销量在 9 月猛增，这并不符合经济的发展规律，正常商家的销售数据，趋势应是较为平稳的，出现这种销量波动剧烈的可能性较低，我们随即对该月、该商家的具体商品信息进行了核查，发现该商家某些商品存在异常刷单的行为。

4.1 基于方差的销量异常检测方法

基于以上分析，我们得出结论：若某商家的销售数据出现较大的波动情况，则说明该商家出现异常刷单行为的概率越大，该商家的异常程度也就越高，基于此我们进行了如下的操作：

4.1.1 通过方差判断异常程度

首先我们将分别去检索每个商家在不同月份的销售情况，并计算总体的方差，随后我们依次去检测每个月的销售数据，在去除本月数据后，计算其他几个月销售数据的方差，如果该月数据对总体方差有较大的影响，则说明该月数据可能存在异常，我们随即去检测该月该商家所售卖的具体的商品信息。

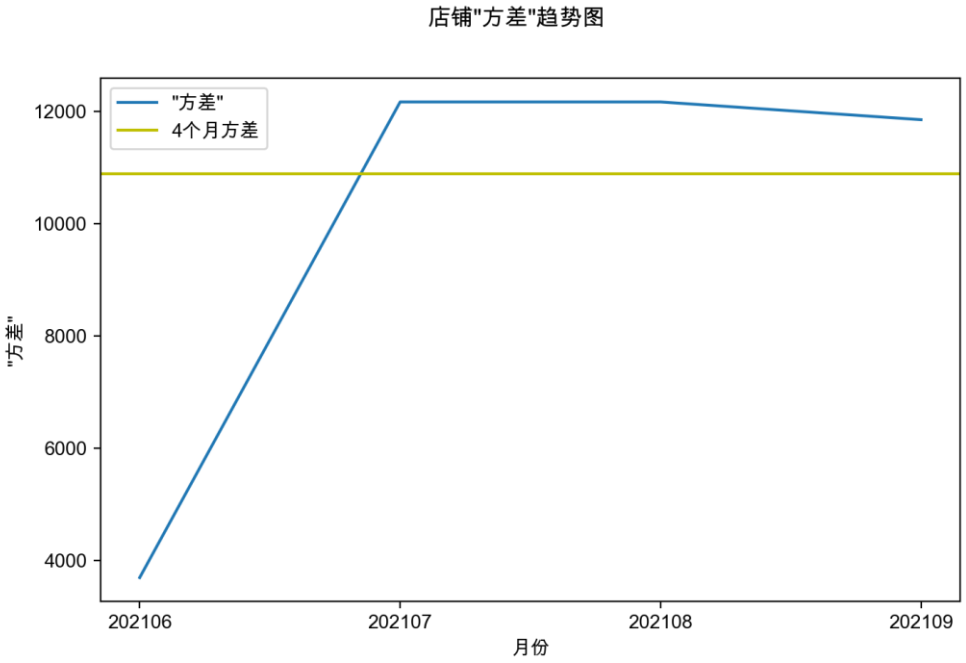


图 4.2 店铺方差趋势图

如上图，可以看出 2021 年 6 月的数据明显低于该店铺四个月平均方差，所以该月数据可能存在异常。

4.1.2 设置合理的销量区间范围

然后我们将该商家所有商品数据对销量进行降序排序，由于销量异常是指销售量偏高，故为了提高模型的性能，我们选取前 30~50%的数据并按照如下方案进行检测：

（1）首先依次遍历每件商品的具体信息，通过它的商品 ID 检索出该商品其他月份的销量情况；

（2）随后我们计算该商品其他几个月销量数据的平均值作为一个基值 X，设定如

下图的区间范围，判断 X 在哪个区间内；

图 4.3 销量区间图

（3）对于不同区间内的数据，我们分别设置了一个变异系数，通过这个变异系数，我们计算出该件商品的销量波动范围，得到一个正常的销量波动的区间，如果本条记录的销量在区间范围内就为正常，否则就为异常。

4.1.3 预测商品销量

考虑到商家可能会出现新品上市的情况，这就导致了该件商品无历史数据进行参考，这时我们就需要对该商品可能出现的销售量，以及变化范围进行预测。因此我们检索出该商品所属类目的其他所有商品的销售数据。

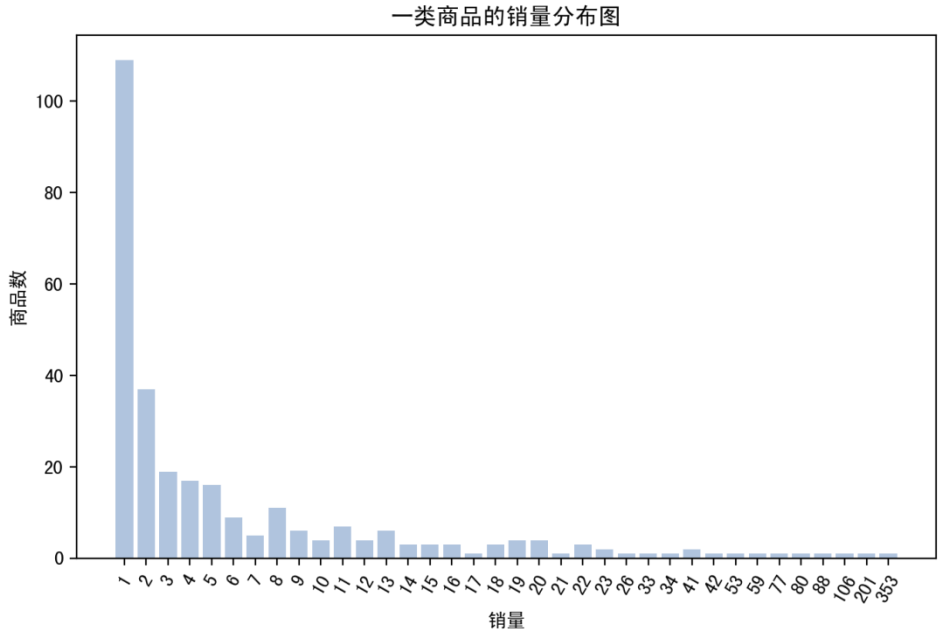


图 4.4 销量分布图

由上图可知，数据符合长尾分布，为了减小极端值对预测的销量产生较大的影响，我们去掉了销量较高的前 5%，以及销量较低的后 2%取中间的 93%的销量数据生成基准值 X 用于预测该件商品的销售情况，后续步骤 4.1.2 中方案相同。

4.2 基于参数评分的销量异常检测方法

（1）参数分析

我们在选购商品时，除了会关注商品的品牌、价格外还会观察该商品的其他参数

信息，如果该商品的参数信息较为完整的话，在同等情况下，用户会倾向于选择此类产品，基于此我们发现参数字段里面包含了大量的产品信息，例如品牌、产地、材质、适用季节等。

（2）计算每个参数中各属性的权重

然后我们对每一类商品进行检测，统计该类商品参数字段中各个属性在整个类别中出现的次数，并定义该属性出现的次数与该类商品总数的比值为该属性的权重，属性出现的次数越多则该属性的权重就越大。

（3）计算评分判断异常值

最后我们分别去检测该类每件商品的参数信息，检测其完整程度如何。评分计算公式如下：

设一个类别有 n 件商品，有属性[a1,a2,…,am]，对应分别出现的次数为[c1,c2,…,cm]，所以可以得到每个属性的权重为

$$w_i = \frac{c_i}{n}$$

对于每一件商品，判断是否包含该类商品中的所有属性，若包含则 bi 为 1，否则为 0，如此可以得到一个 0,1 向量[b1,b2,…,bm]，设其中 1 的数量为 x，则评分为

$$S = \frac{b_i w_i}{x}$$

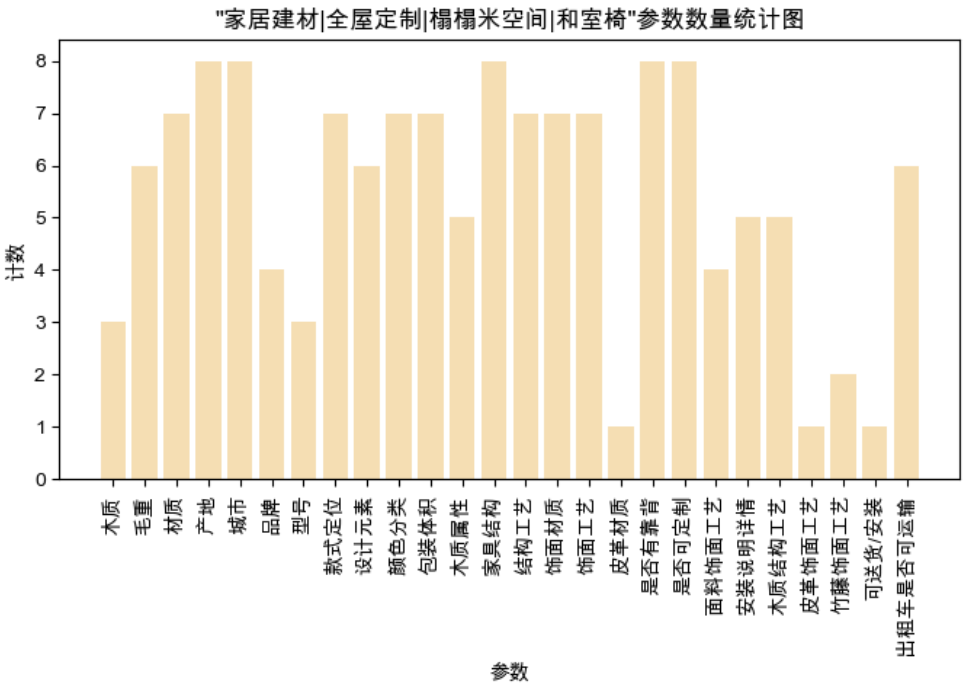


图 4.4 各类参数数量统计图

分析上图可以看出，皮革材质、型号等参数数量都十分少，可以认为不是特别重要；而材质、是否有靠背、家具结构等参数数量很多，说明该参数权重更大，因此包

含改参数的商品对应的评分就越高，如果该字段属性评分越高，则代表该商品的可信程度越高，异常程度越低。

对所有商品数据都进行上述操作即可计算出其对应的评分，其中评分低于 0.5 的我们认为是异常数据。