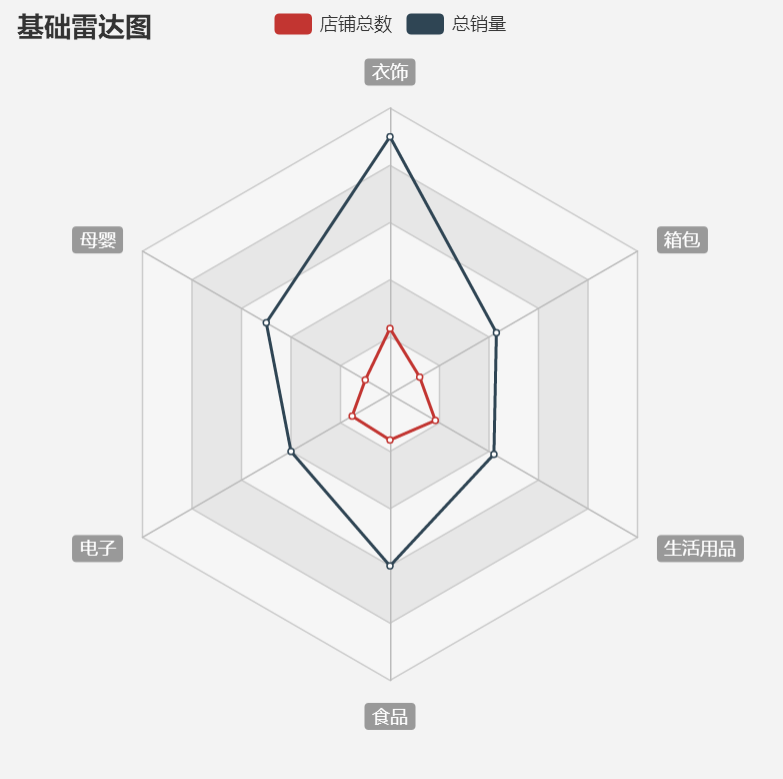
**基于淘宝网商家信息的数据分析与可视化**

刘晓 李浥菡 张帆

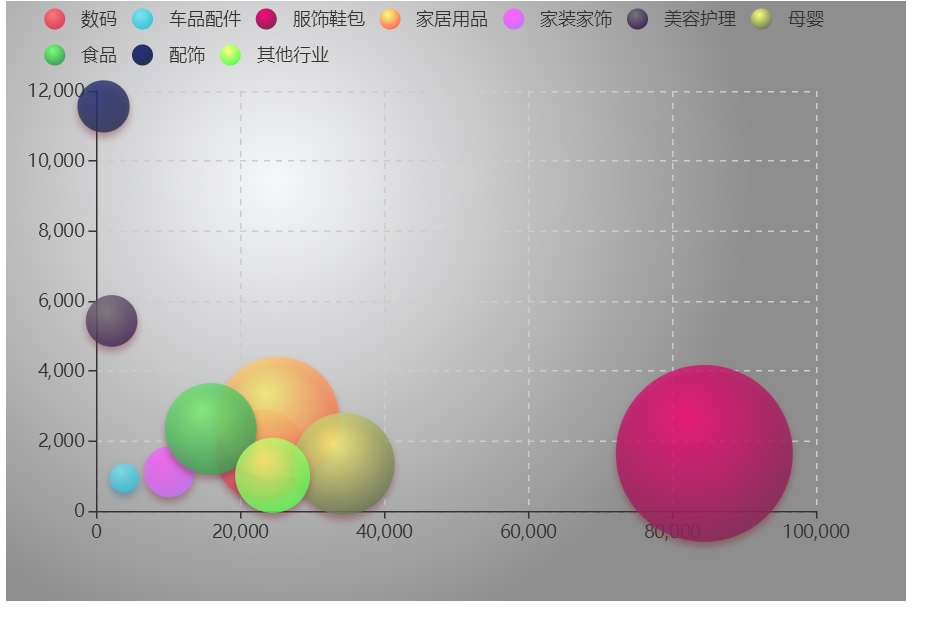
1. 主要内容

**1.1各类店铺总销量的分析与可视化**

从衣饰、箱包、生活用品、食品、电子、母婴六项品类入手，通过数据抓取，分析各类店铺总数及该类店铺总销量的关系，利用可视化方式，呈现出各类店铺的发展程度、饱和程度，进而分析各类店铺的供求关系。已入驻商家可参考该数据，进行品类调整；并可向新入驻商家提供品类选择方案。

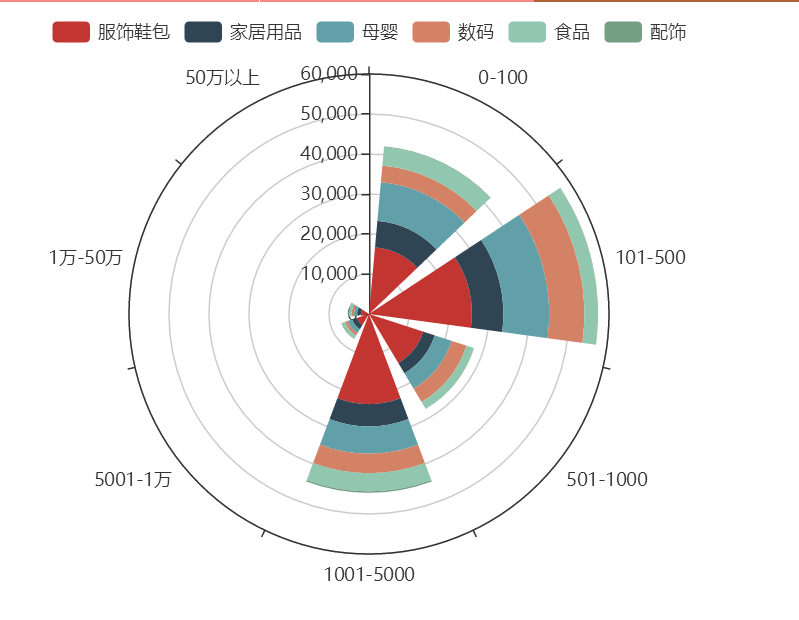


**如下图所示，以圆的半径大小代表该类商品的店铺总销量，坐标轴横坐标代表该类商品的店铺总数，纵坐标代表该类商品的店铺平均销量。不难看出服饰鞋包类商品的店铺虽然总销售量很高，但同时由于此类店铺总量较多，因此该类商品的店铺平均销量不是很高；与之形成对比的如配饰类商品，此类商品的店铺总销量远不及服饰鞋包类商品，但因其店铺数量较少，因而该类店铺的平均销量比较可观，平均可达10,000件以上。据此在引导新商家入驻平台时时，可适当倾向于如配饰类、美容护理类商品的销售，这样既可以丰富平台商品的种类，也利于消除各类店铺间发展的差距，实现商家与平台的互利双赢。**



**1.2同类商品的不同销量阶段店铺数量的分析与可视化**

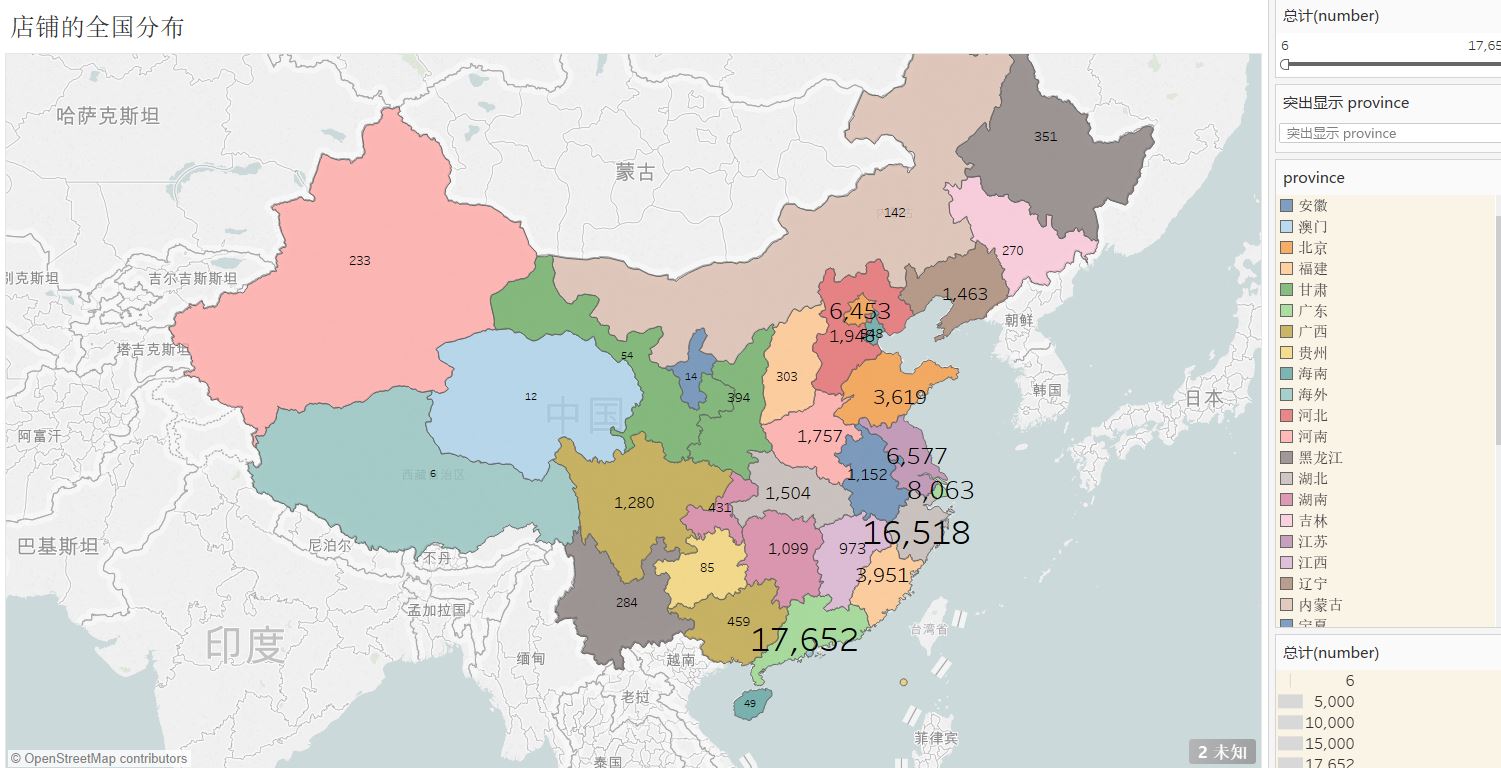
通过对同类商品的不同价格档位的销量阶段（0—500,500—1500,1500—3000，3000—5000,5000—10000，10000以上）店铺数量所占的百分比，可分析出该品类下店铺的经营状况、水平及定位，以及因为一些不可控制因素导致的过低或者过高销量（例如玉器、金银首饰等比较贵重的东西可能导致销量过少等等），进而分析出消费者的消费水平、喜好，商家可以参考数据合理定价或进行价格调整。

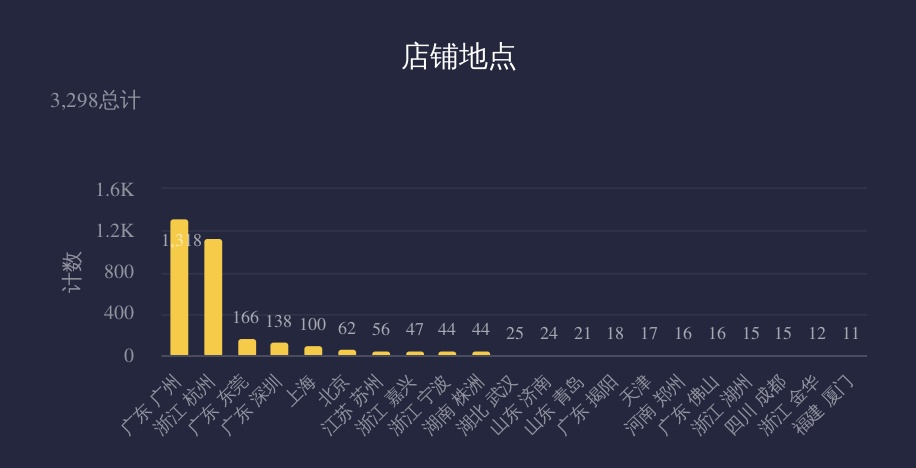


**1.3店铺地址在全国分布区域的分析与可视化**

通过统计店铺地址在全国的区域分布，可体现淘宝店铺在全国的分布率，由此可分析淘宝在各区域的活跃程度。通过分析同类店铺在全国的区域分布，可看出该类商品在的该区域的发展成熟度（例如广东的衣服工厂）；通过分析各区域集中发展的品类，可分析该品类的在当地的生产成本、人工成本等，对商家选址、选项具有一定指导作用和参考价值。

**如下图所示，我国东部沿海地区淘宝店铺数量明显高于中西部地区，其中以珠三角及长三角地区最为突出，店铺地址位于广东省内的商家最高可达17,652个，江浙沪三地紧随其后，从数量上不难看出这些省市的店铺发展较为完善，更具市场潜力。分析表明珠三角地区的服装加工产业与江浙沪地区的小商品制造业极具优势，这两类产品已初步形成产业群独立发展，同时考虑到东南沿海地区的廉价劳动力优势、国家政策扶持、销售渠道多元化等一些列优势，在商家进行店铺选址及选项时，可适当参考本地区的店铺总量及发展现状，以对店铺的未来发展理性预期。**

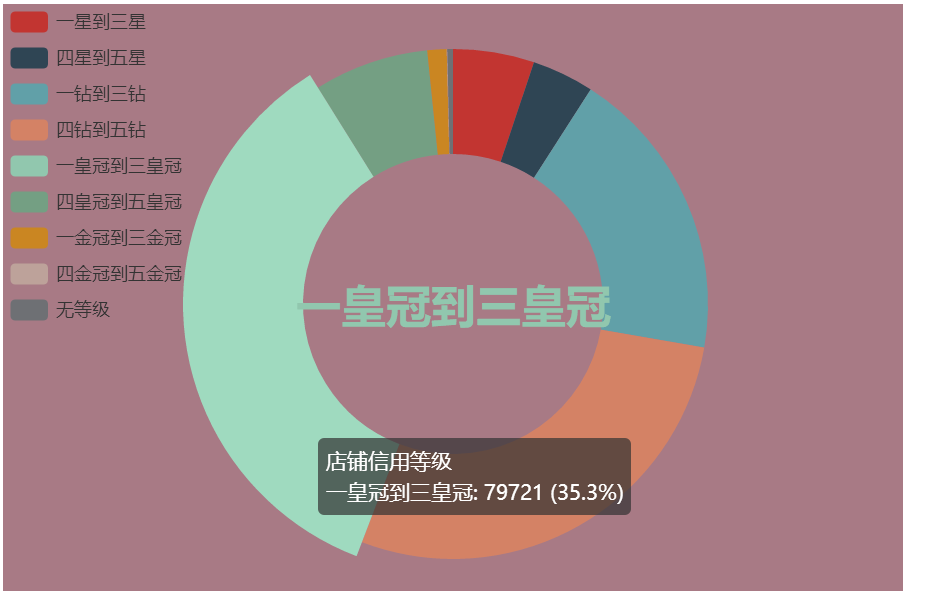




**1.4淘宝店铺生存状态的分析与可视化**

根据淘宝店铺等级以及所占店铺总数的比例，可视化呈现出淘宝店铺的生存状态。此项数据也反映出淘宝店铺大部分处于一种低迷的经营状态，入驻门槛低，但发展情况良莠不齐，想要发展为优质商家仍需从商品品质、营销手段、价格定位、用户体验等众多方面进行提升。

下图是店铺星级分布图：



下图是积分等级分布图：



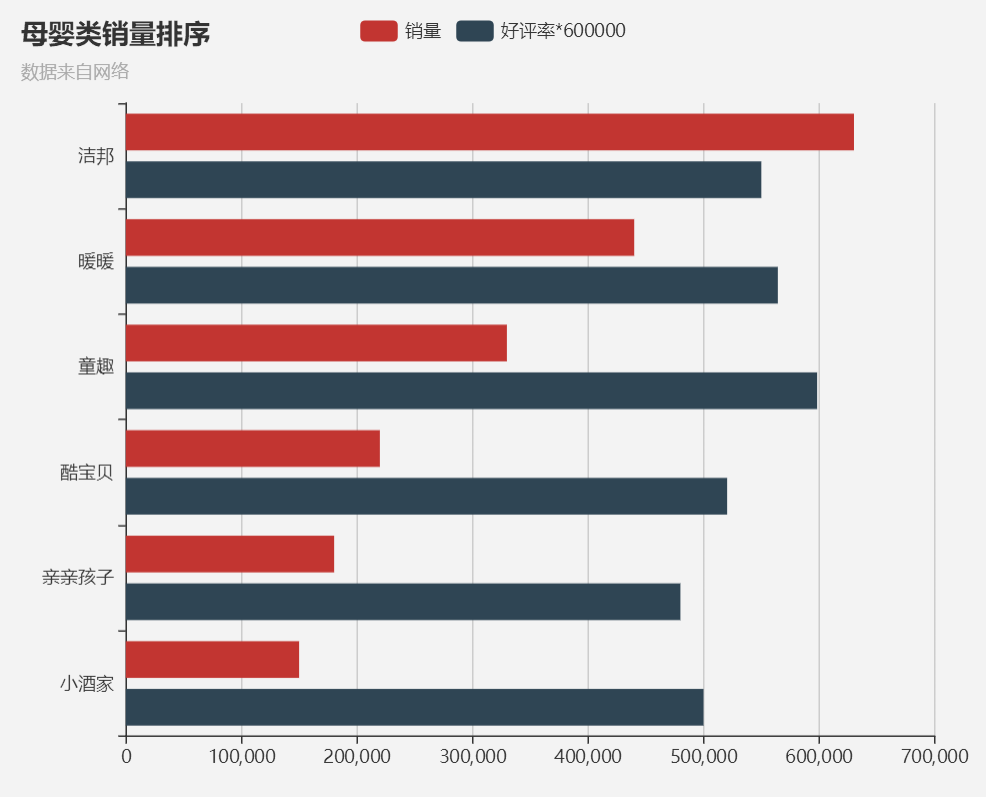
**目前淘宝店铺信用等级的分布差别较大，占最大比重的属一皇冠到三皇冠的商家店铺，占到总量的35.5%，此类店铺当属销售中的佼佼者，从信用等级和积分等级来看都是极具竞争优势的；其次四钻到五钻的店铺也紧随其后，此类店铺具有较大发展潜力，约占总量的三成；而一钻到三钻的店铺达到两成左右，其余等级的商家数量较少，约总量的四分之一。**

**淘宝店铺的评分主要来自于店铺信用等级和交易积分等级，因此如何提高商品质量、提交成交量、加大宣传等一系列能带动客户体验的举措皆是影响店铺发展的关键因素。**

**1.5淘宝店铺性价比的分析与可视化**

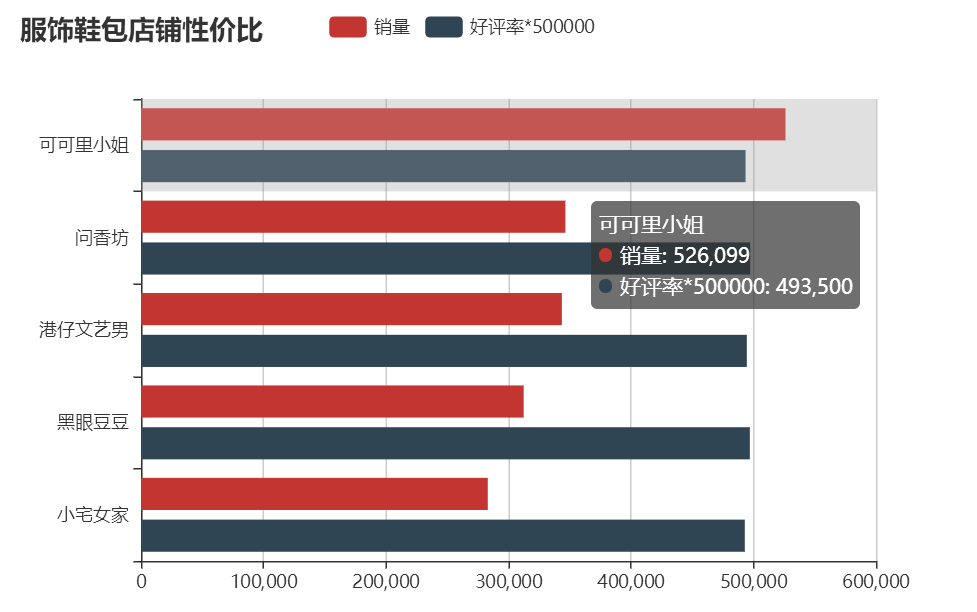
根据每品类店铺销量进行排序，结合其好评率综合分析，选出每类商品性价比较高的店铺（排名前5家），并可向消费者定向推荐。

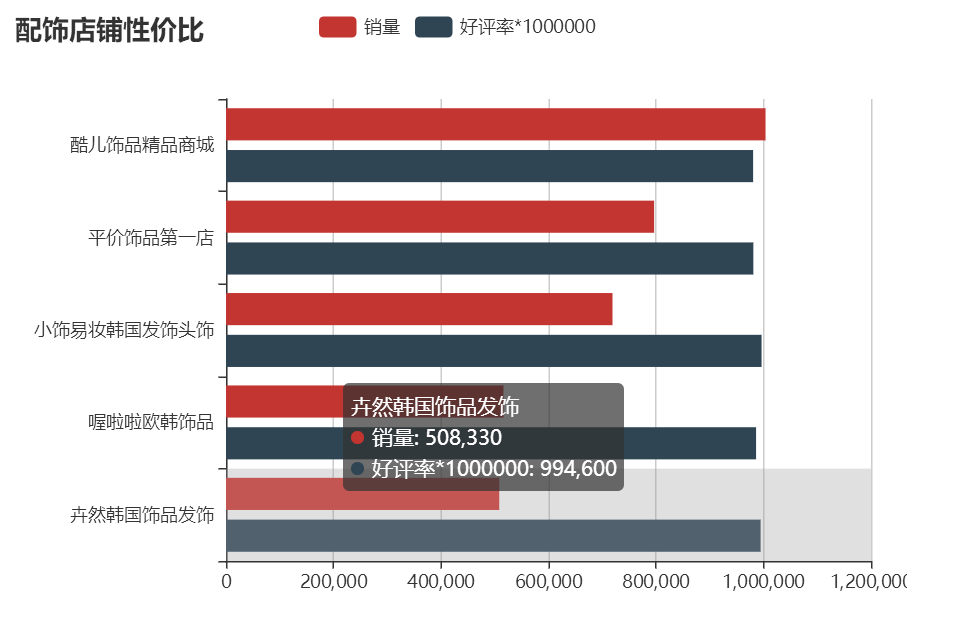
**根据淘宝平台上各类别店铺的销量进行排序，同时综合商品的好评率进行分析，筛选出此类别商品性价比较高的店铺，可定期向消费者进行定向推荐。此举可规范商家的销售行为，给“优质”商家起到推广宣传作用，给“普通”商家起到引导示范作用，同时也为消费者提供便捷、可信的选择渠道，合理进行网购。**

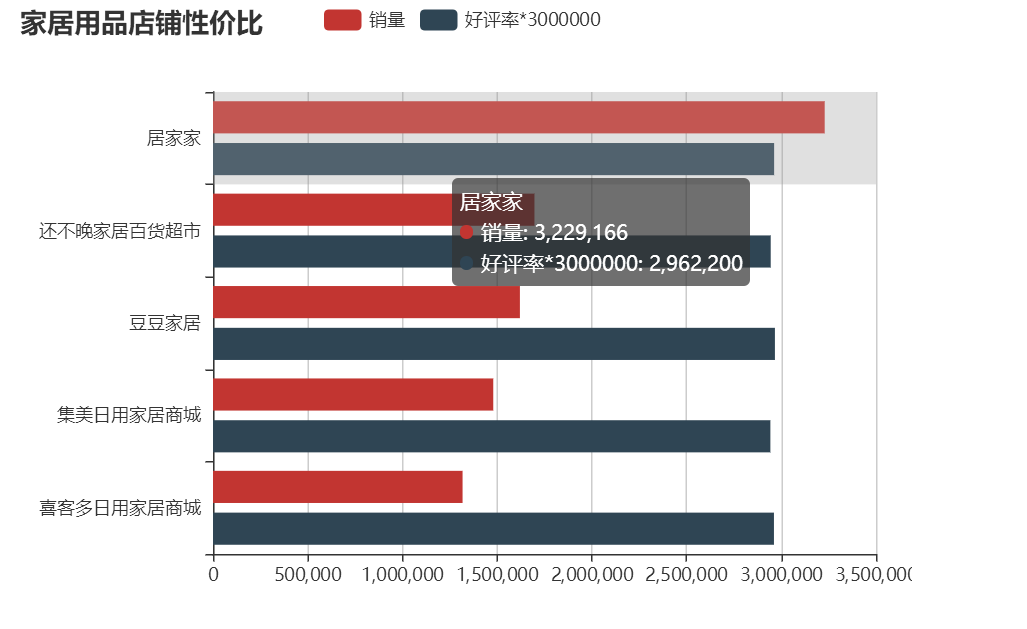


**1.6商品评论中好评率相近销量排名前五店铺的分析**

根据5得出的性价比分析结论，进而收集其店内商品评价内容的情感词词云，对评价内容中的情感词汇做出情感分析，可将其分为积极情感、消极情感、中立情感；通过分析、筛选，进而区分有价值信息及无效信息，有选择性、针对性地将商品评价内容转化为词云，更加直观地体现出用户更关注的方面，并对消费者的购物选择具有导向作用。（销量排名前三的店铺）







**1.7根据商品评论进行年度分析**

**将根据5得出的性价比高的店铺中评价内容进行年度分析，出具数据整合报告，高频词的出现和变化体现出此店铺商品的整体变化趋势，以多种维度来综合考量，定期对此店铺的商品质量及经营状况进行理性分析，并能够引导店铺及时调整，适应市场发展。**

**以下是服饰、电子设备、配饰、居家用品四大类中销量最高的四个店铺。**

（1）vero moda:连衣裙（该类下销量最高）



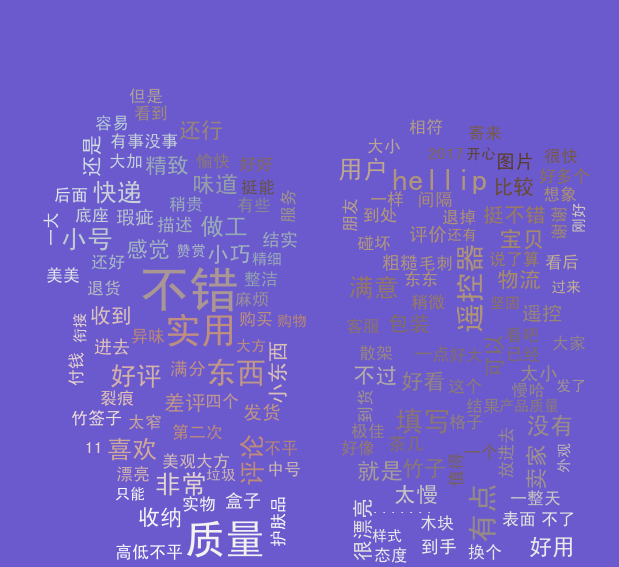
评论内容：<https://rate.tmall.com/list_detail_rate.htm?itemId=545358504014&spuId=722589706&sellerId=420567757&order=3&currentPage=2>



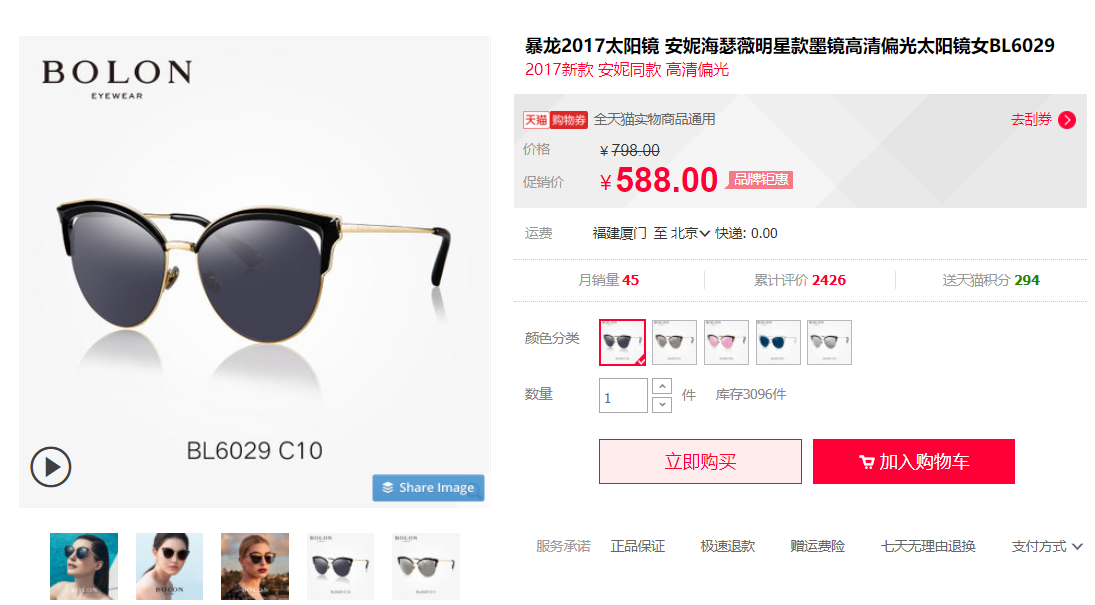
（2）漫步者宏通旗舰店：低音炮



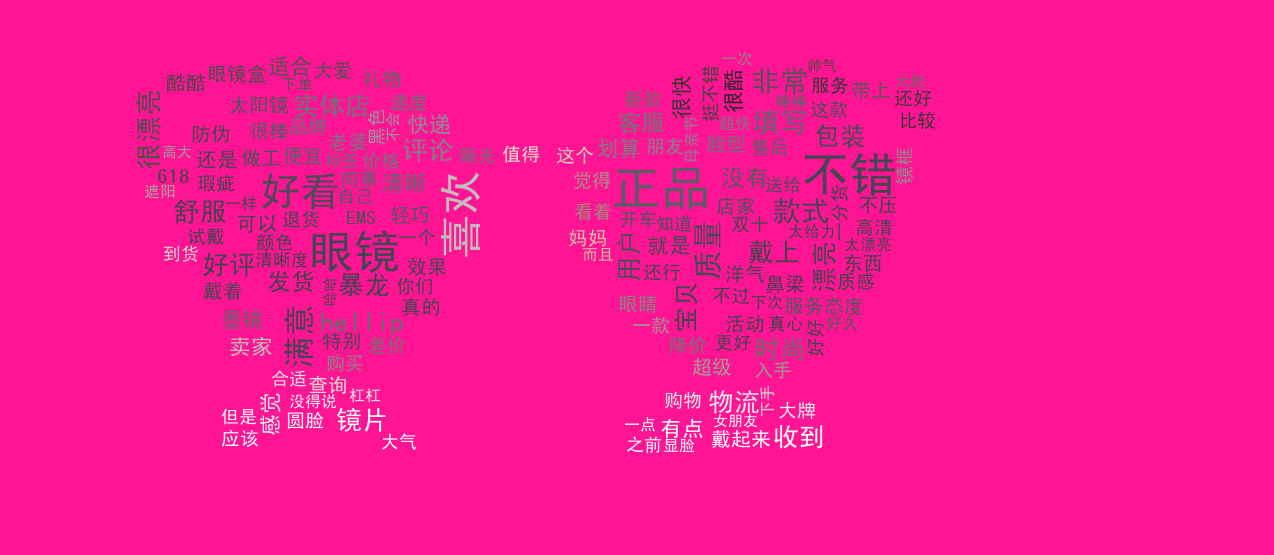
评论内容： [https://rate.tmall.com/list\_detail\_rate.htm?itemId=520491114368&spuId=339731168&sellerId=2098207494&order=3&callback=jsonp1873&currentPage=30](%20https:/rate.tmall.com/list_detail_rate.htm?itemId=520491114368&spuId=339731168&sellerId=2098207494&order=3&callback=jsonp1873&currentPage=30)



（3）暴龙官方旗舰店：太阳镜



评论内容：<https://rate.tmall.com/list_detail_rate.htm?itemId=544180188944&spuId=719194035&sellerId=374632943&order=3&currentPage=2>



（4）居家家旗舰店：收纳盒

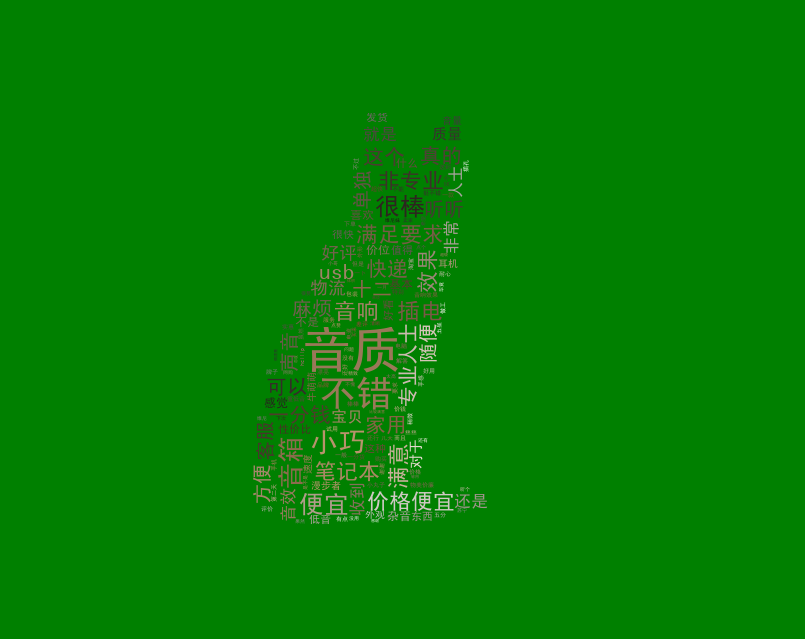


评论内容：<https://rate.tmall.com/list_detail_rate.htm?itemId=536483439040&spuId=689371740&sellerId=2938454519&order=3&currentPage=2>



纵向来看同一产品2017年和2018售后评价内容情况，以此来分析该商品在2017-2018年间质量是否有所变化：（以下为对商品低音炮的分析）

**2018年评价：**



**2017年评价：**



1. **结论**

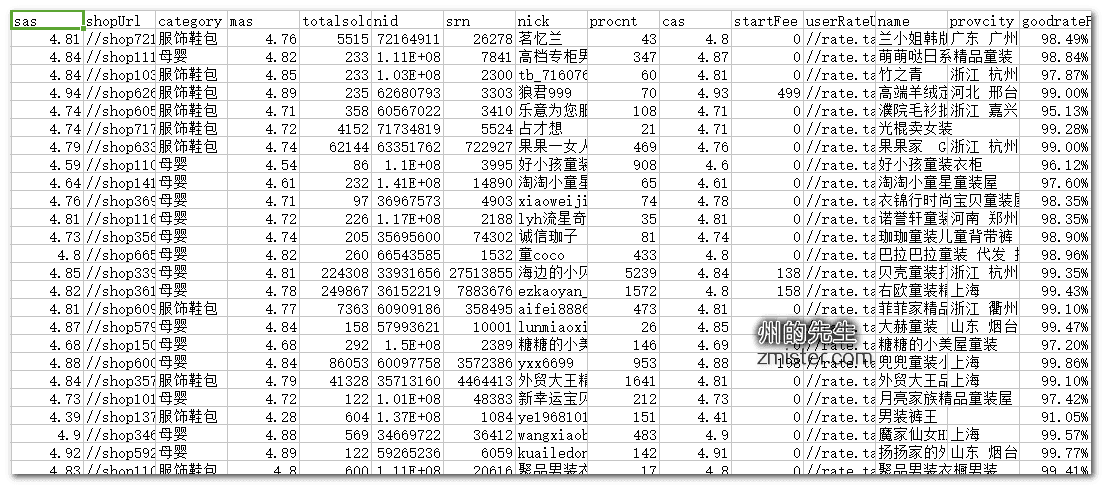
通过对淘宝网商家的信息进行汇总分析，初步得到结论包括以下几点：目前各类别的淘宝商铺经营状况差距较大，各地区商铺发展水平不均衡，鉴于平台对商铺的评分机制，商家应通过多种渠道来增加交易量，提高店铺等级；同时买方对商品的认可和评价也尤为重要，它对店铺及此类商品在平台的供求关系起一定作用。

现实中淘宝商铺经营的因素是多种多样的，只有通过多角度、深层次的数据分析才能窥探其中的奥义，不断地以数据分析的方式去调整、去完善经营策略、监管机制，才能实现买家、商家及平台共赢的目标。

1. **技术路线**

**2.1数据来源与获取方法**

使用Python爬虫获取淘宝网上（指定数量）的商家信息，包括店铺名称，店铺区域，店铺销量，店铺好评率，店铺分类，店铺等级、以及部分商品的评论等信息。



**2.2分析与统计**

统计与分析指定数量的淘宝商家信息中，不同类型的店铺数量，同类别店铺发货地的分布区域，同类别商品评论内容的高频词，每个店铺的销售量，不同等级的店铺数量。

**2.3可视化与展示**

同类商品的不同销量阶段店铺数量所占的百分比

[http://echarts.baidu.com/demo.html#bar-polar-stack-radial](http://echarts.baidu.com/demo.html" \l "bar-polar-stack-radial)

**2.4可能会遇到的问题与解决办法**

（1）淘宝数据过多且爬取过多会使得访问受限，故只爬取部分商品评论，商家信息数据来源网络获取。

（2）分析数据包，找到淘宝评论传输用的网址，分析评论内容的特点。

（3）如何从爬到的json格式数据中得到想要的数据，进行分析与可视化。

1. **进度安排**

12.25-12.28 爬取数据及初步分析

12.29-01.08 对数据进行可视化

01.08-01.18 搭建可视化网站，完成项目报告