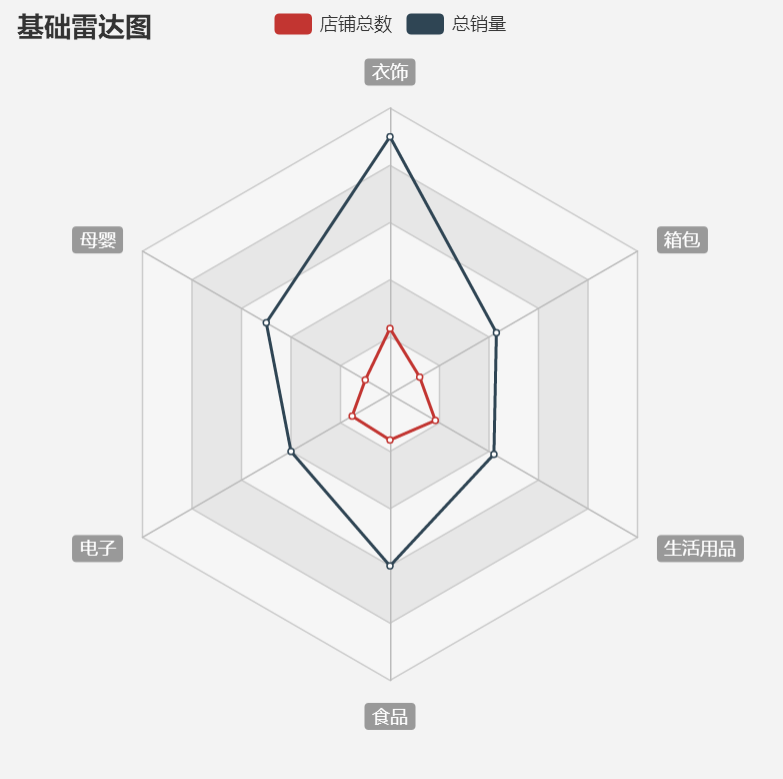
**基于淘宝网商家信息的数据分析与可视化**

刘晓 李浥菡 张帆

1. **主要内容**

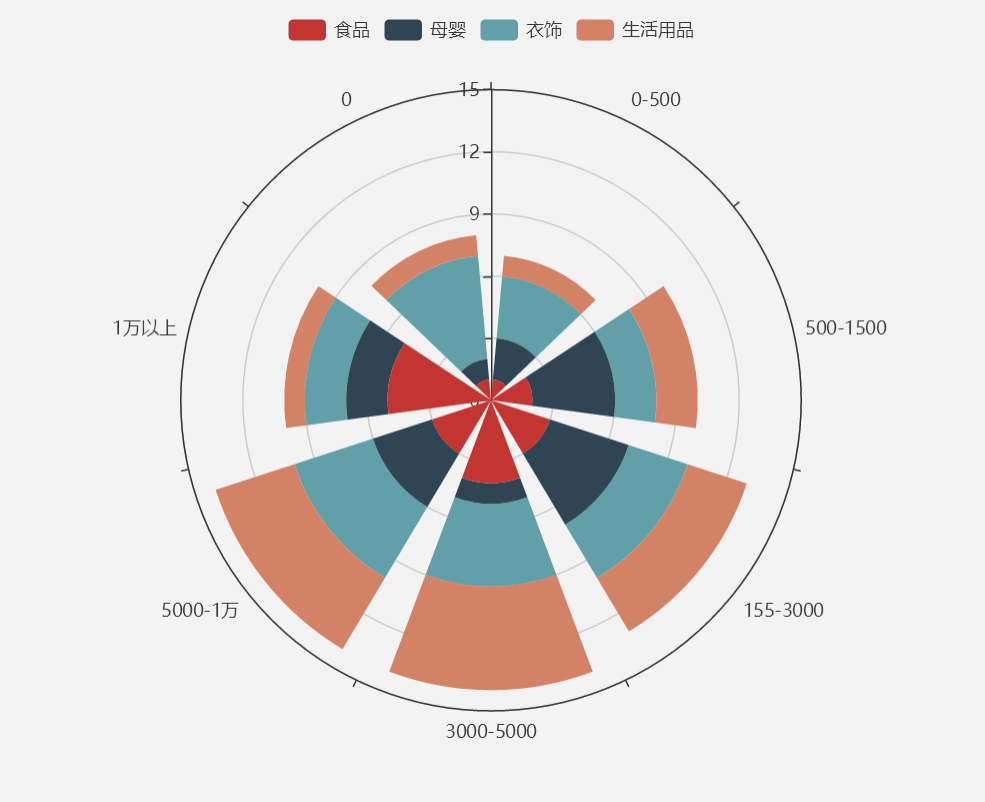
**1.1各类店铺总销量的分析与可视化**

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。



**1.2同类商品的不同销量阶段店铺数量的分析与可视化**

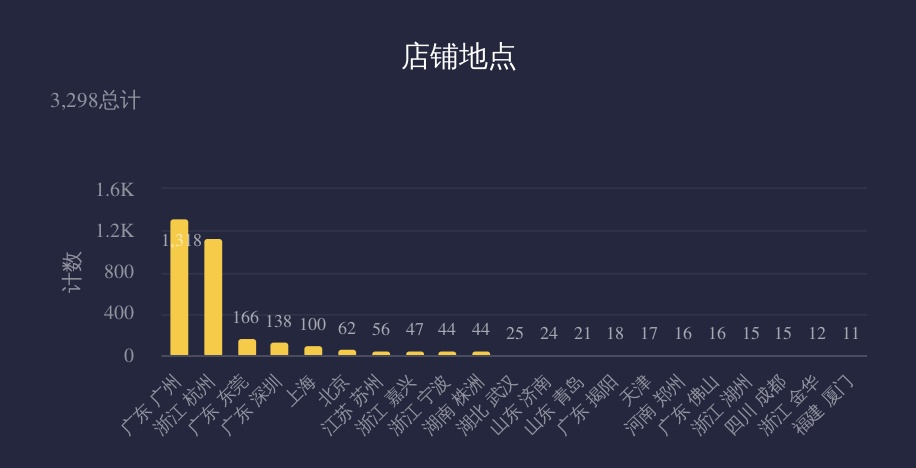
通过对同类商品的不同价格档位的销量阶段（0—500,500—1500,1500—3000，3000—5000,5000—10000，10000以上）店铺数量所占的百分比，可分析出该品类下店铺的经营状况、水平及定位，以及因为一些不可控制因素导致的过低或者过高销量（例如玉器、金银首饰等比较贵重的东西可能导致销量过少等等），进而分析出消费者的消费水平、喜好，商家可以参考数据合理定价或进行价格调整。



**1.3店铺地址在全国分布区域的分析与可视化**

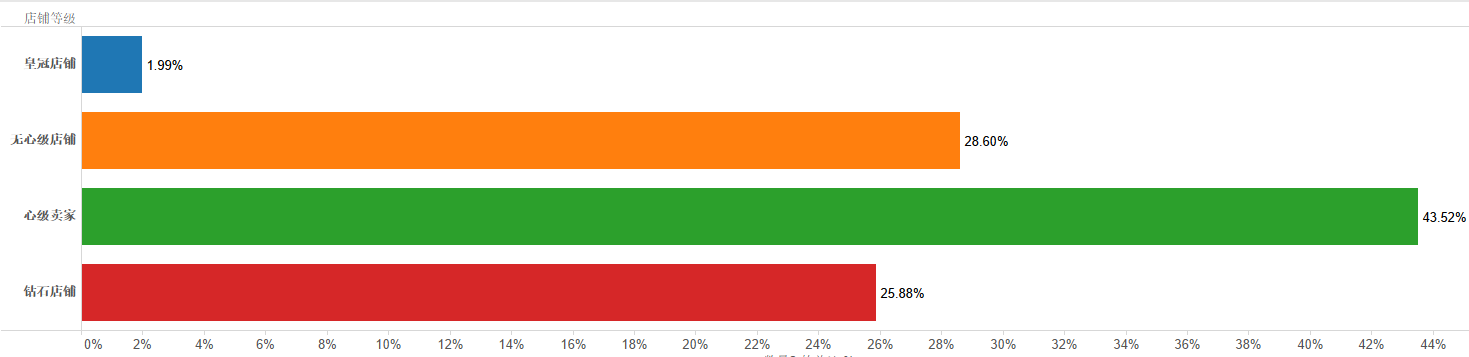
通过统计店铺地址在全国的区域分布，可体现淘宝店铺在全国的分布率，由此可分析淘宝在各区域的活跃程度。通过分析同类店铺在全国的区域分布，可看出该类商品在的该区域的发展成熟度（例如广东的衣服工厂）；通过分析各区域集中发展的品类，可分析该品类的在当地的生产成本、人工成本等，对商家选址、选项具有一定指导作用和参考价值。





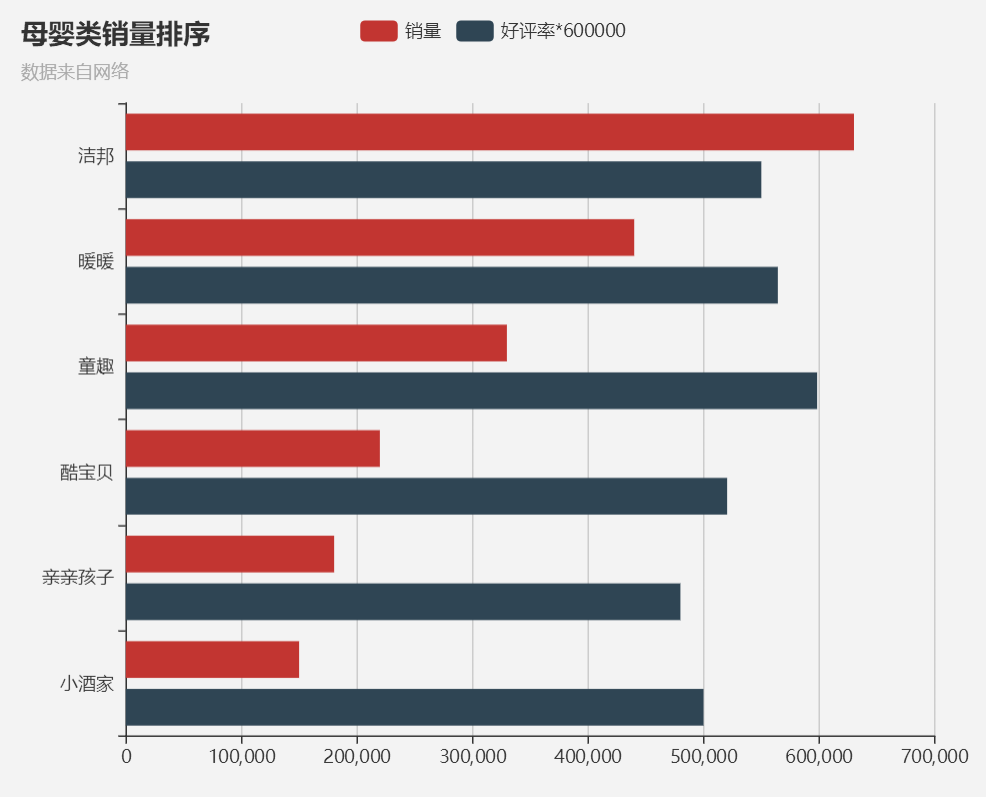
**1.4淘宝店铺生存状态的分析与可视化**

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。

****

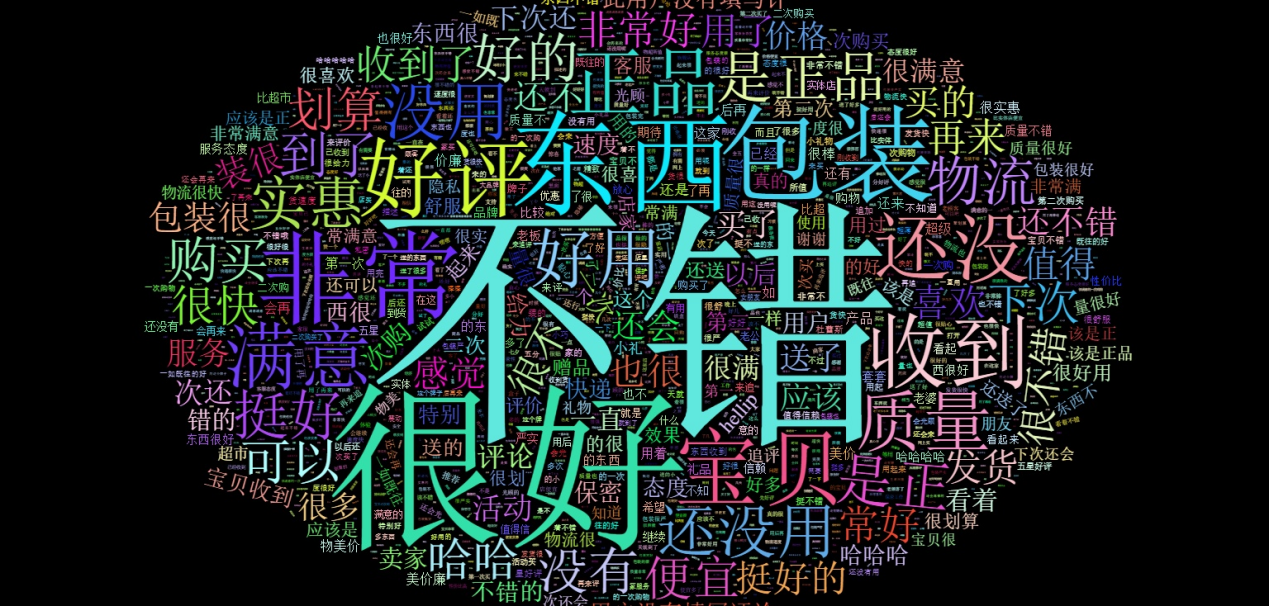
**1.5淘宝店铺性价比的分析与可视化**

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。



**1.6商品评论中高频词的分析与可视化**

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。



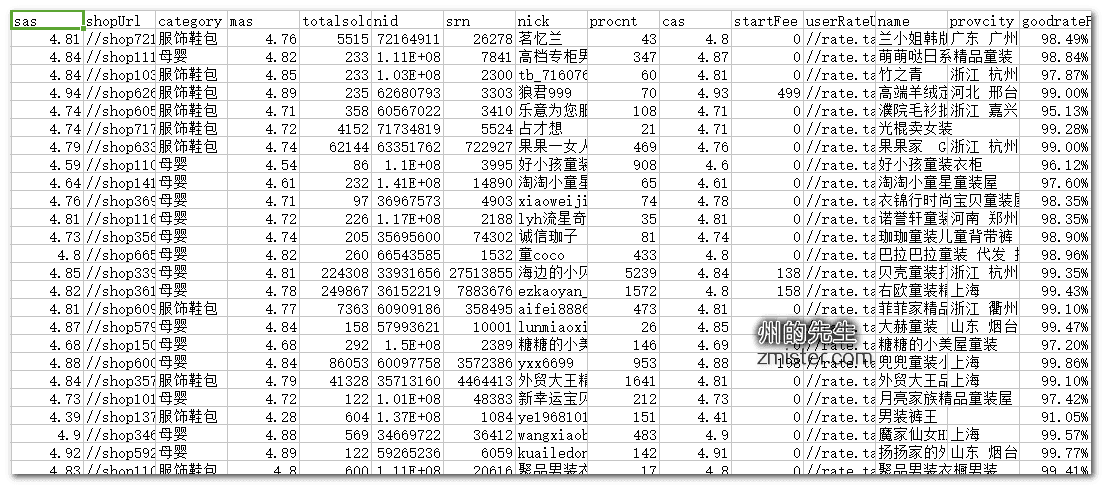
**1.7根据商品评论进行年度分析**

将根据5得出的性价比高的店铺中的评价内容进行年度分析，出具数据报告，可体现出2016年和2017年评价内容，分析近两年该店铺商品的质量状况。

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、商品评论等信息。



**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，故只爬取部分数据。

（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

12.29-12.31 对数据进行可视化

01.01-01.08 搭建可视化网站，完成项目报告