

数据分析报告

By 喵喵喵

队长：121250101 缪晓伟

队员：121250169 吴雨婷

121250135 孙艺博

目录

一、 数据源说明	1
二、 统计分析方法	2
1. 商品价格走势与价格分布	2
2. 商品评论数统计	2
3. 星级评分统计	2
4. 分析节日对销量的影响	2
5. 分析评论长度与评分的关系	3
6. 分析价格走势与评论数量增量的关系	3
7. 分析不同卖家出售的同一商品的价格走势	3
8. 出售或某一类别商品的商家数量	3
三、 结果图表与分析	4
1. 商品价格走势与价格分布	4
2. 商品评论数统计	5
3. 星级评分统计	5
4. 分析节日对销量的影响	6
5. 分析评论长度与评分的关系	6
6. 分析价格走势与评论数量增量的关系	7
7. 分析不同卖家出售的同一商品的价格走势	8
8. 出售或某一类别商品的商家数量	8
四、 多余的话	10

一、数据源说明

此次数据分析使用的所有数据均来自 `trendada.cn`，通过网站开放的 Amazon REST API，使用 Python 语言写成的脚本来获取不同的分析功能所需的不同数据。由于对外展示的前台尚未完成，获取到的数据在经过适当的格式化后，手动填充入生成相应的 Highcharts 图表的 JavaScript 脚本中。

目前，所有直接通过 API 获取数据的方法已经完成。各个方法没有获取 API 能够提供的所有内容，而是分别按需求选择部分域获取，以此提高 API 的访问效率，同时尽可能减轻数据获取给服务器带来的压力。

我组在数据分析时选择了四个商品分类，清单如下：

- Clothing & Accessories>Baby>Baby Girls
- Clothing & Accessories>Baby>Unisex
- Health & Personal Care>Household Supplies>Air Fresheners
- Health & Personal Care>Household Supplies>Household Cleaning

这四个分类分别包含数百件不同商品，可以满足此次数据分析对商品规模的需求。

此报告中所有图表、文字生成于 2014 年 4 月 27 日，展现的数据均为 REST API 于 2014 年 4 月 27 日提供的数据。生成结果的代码在 Python Shell 中直接键入，并没有作为永久文件保存的打算，故在此恕不提供。

为了简单起见，商品销售的商户中，这里直接将图片作为商户识别标志，商户名为 'Null' 的商户排除在外，不予统计。

二、统计分析方法

除三个在要求中声明必须实现的功能外，我组另外选择实现了以下五个功能：

- 分析节日对销量的影响；
- 分析评论长度与评分的关系；
- 分析价格走势与评论数量增量的关系；
- 不同卖家出售的同一商品的价格走势；
- 出售过某一类别商品的商家数量。

下面阐述我组对上述各个功能的展示方法。

1. 商品价格走势与价格分布

对单个商品的价格走势，将通过折线图展现出来。由于顾客往往只关心该商品在某个时间点的最低价格，因此若同一时间（数据爬取时间）有多个商家出售同一件商品，我们选择只记录最低的价格。同样地，在统计某一类商品的价格分布时，我们默认对商品的最新、最低价格进行统计和汇总。由于价格的分布结果展现是对一个分类下所有商品价格的罗列，在本报告中展示不甚美观，这里不再展示。此功能的演示将在最终产品中提供。

2. 商品评论数统计

由于每一个分类中的商品数量较多，对一个分类中所有商品的评论数比较在一张图表中展现过于拥挤，此报告中将展现不同分类中所有商品的评论数之和。通过对商品评论数的统计，可以从中容易地辨别出哪些分类的商品、哪些单件商品是热门的。

3. 星级评分统计

单个商品的星级评分统计将通过柱状图展现出来。由于商品的差异，将分类下所有商品的星级评分累加展现没有意义。通过星级评分统计，可以从一个侧面判断商品的质量和受欢迎程度。

4. 分析节日对销量的影响

此功能旨在分析用户购物习惯，是否在重大节日附近有更多的消费

行为。这里通过对一件商品在不同时间的销售数量的折线图展示，以期能直观地展现出这一趋势。由于评论与实际的购买行为在时间上有一定的滞后，因此在分析判断时对“节日”的时间定义需要放宽。

5. 分析评论长度与评分的关系

此功能旨在分析用户心理，对商品怎样的喜好程度会令用户写出更长的评论。这里通过将一件商品所有评论的星级和评论长度（以单词计）以散点的方式展现在图上，以期能直观地观察出其中的关系。至于观察产生的推断是否合理可信，这将在学习推断统计之后通过更加严密的方法加以证明。

6. 分析价格走势与评论数量增量的关系

此功能旨在分析用户心理，商品价格的变化是否会影响销量。通过将销量和评论数增量的折线图展现在同一张图上，我们能够直观地观察出其中的关系。

7. 分析不同卖家出售的同一商品的价格走势

此功能旨在帮助商家判断自身对某一件商品的定价水平，以及及时做出调价的策略。这里通过对单件商品所有的售价按商家绘制不同的折线，展现不同商家的价格趋势，以帮助商家获取明确的判断：我的商品卖贵了吗？如果降价，该降低到什么水平？

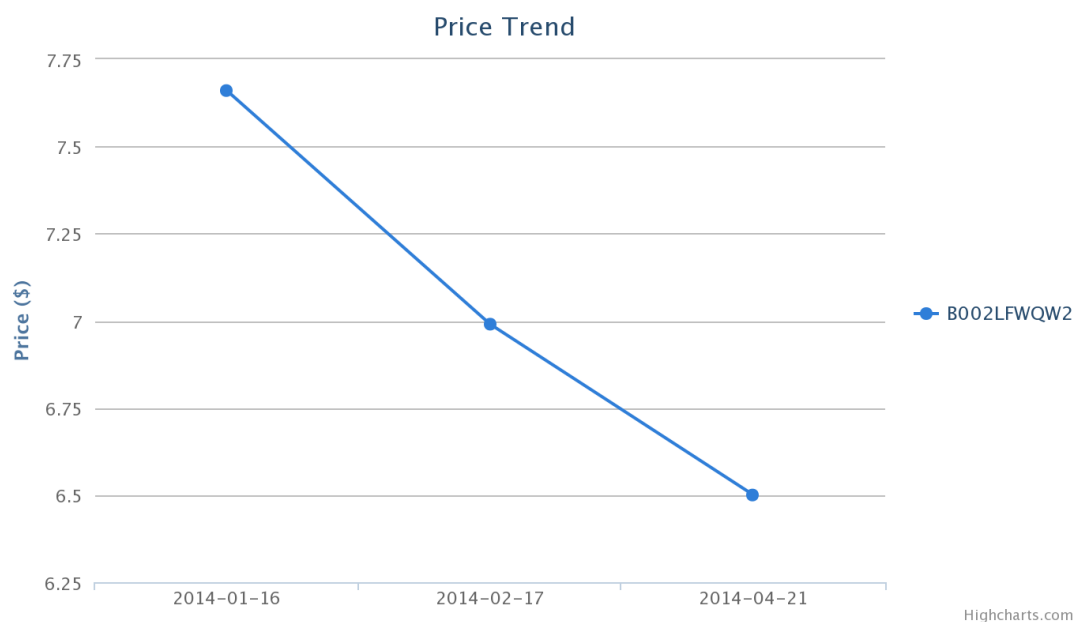
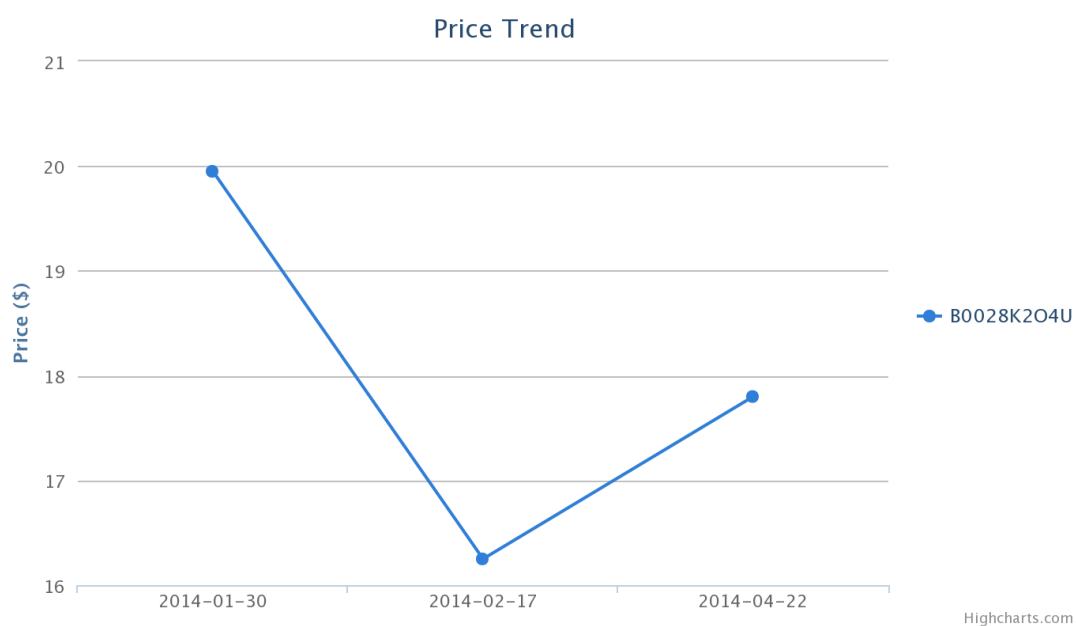
8. 出售或某一类别商品的商家数量

此功能旨在预测热门的商品门类。一件商品或是一类商品的热门与否，除了销量以外，出售的商家的数量也能对此有一定的反映。这里通过对选定的四个分类入驻商家的数量通过柱状图进行比较，以判断出相对热门的商品分类是哪一个。

三、结果图表与分析

1. 商品价格走势与价格分布

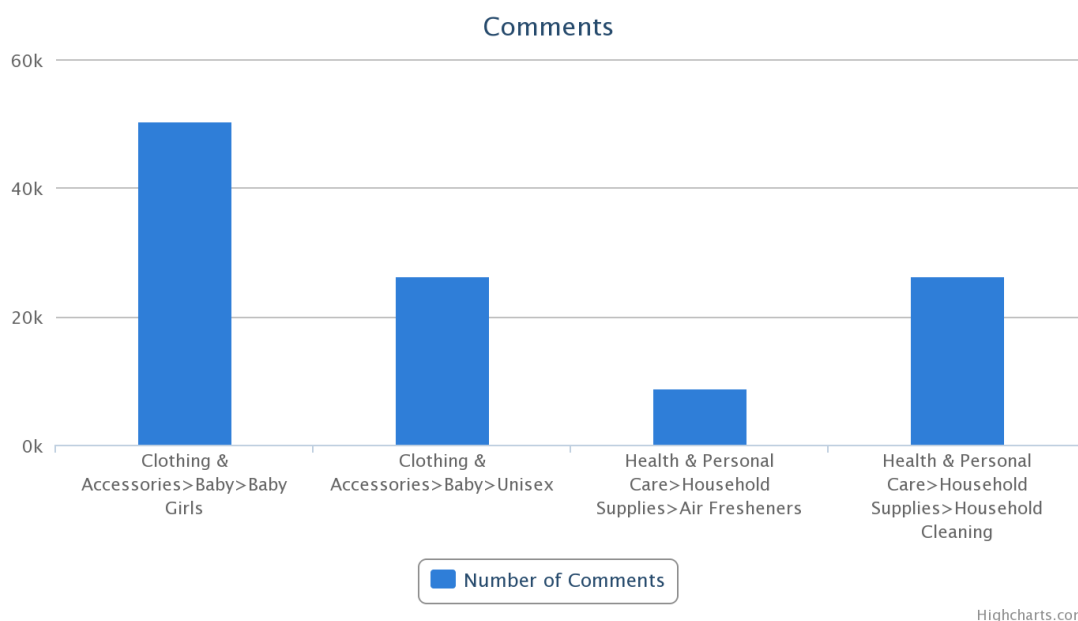
这里以 ASIN 码为 B0028K2O4U 和 B002LFWQW2 的商品为例，展示它们的价格走势。



可以看出，商品 B0028K2O4U 在今年价格有起伏，而商品 B002LFWQW2 的价格一直走低。

2. 商品评论数统计

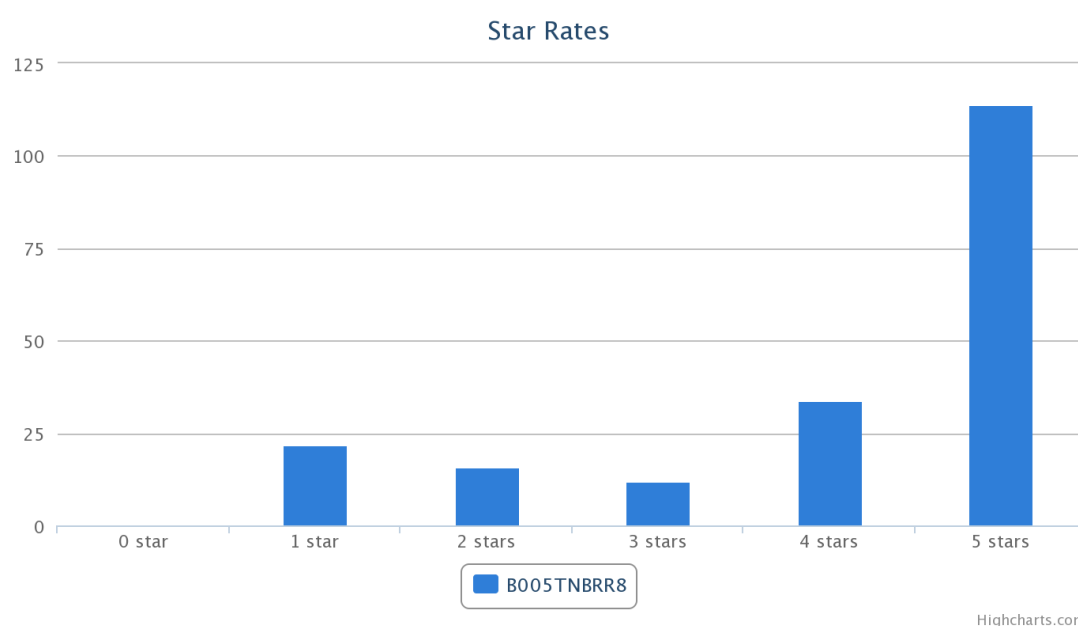
这里统计选定的四个分类的商品评论总数。



从图中可以看出，女童童装的评论数量更多，销量也更好。此图中将空气清新剂和家具清洁工具拿来与服装做比较可能并不合理，这里仅作为展示使用。

3. 星级评分统计

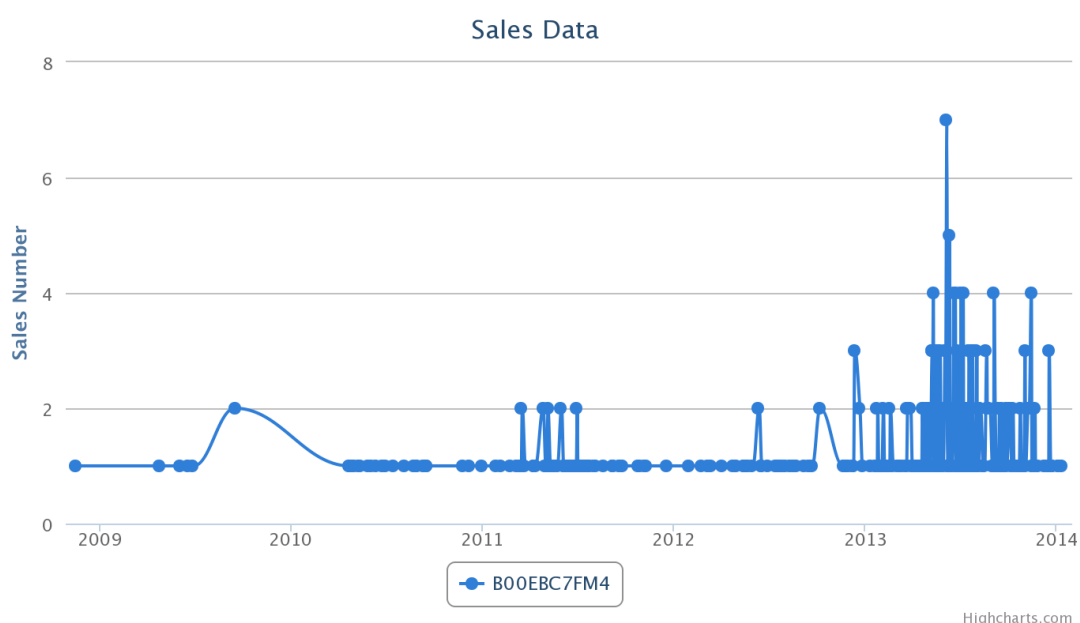
这里对 ASIN 为 B005TNBRR8 的商品进行评论星级的统计。



从图中可以看出，此商品以 5 星评价为主，但负面的 1 星评价也较多。购买此商品时需要对这些低星级的评价仔细参考后再慎重做出购买与否的决定。

4. 分析节日对销量的影响

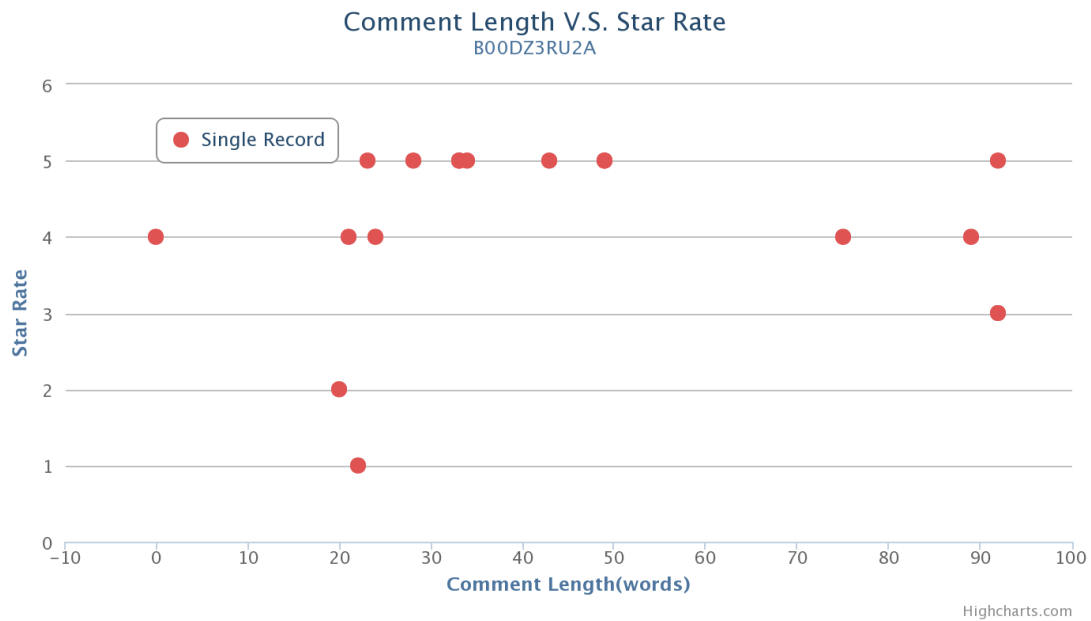
这里对 ASIN 为的商品的销售时间进行分析。由于报告页面宽度限制，图表在此显示过于拥挤，敬请见谅。



从图表中可以看出，该商品的大量销售集中于每年复活节期间，这说明节日对商品的销售有一定的影响。

5. 分析评论长度与评分的关系

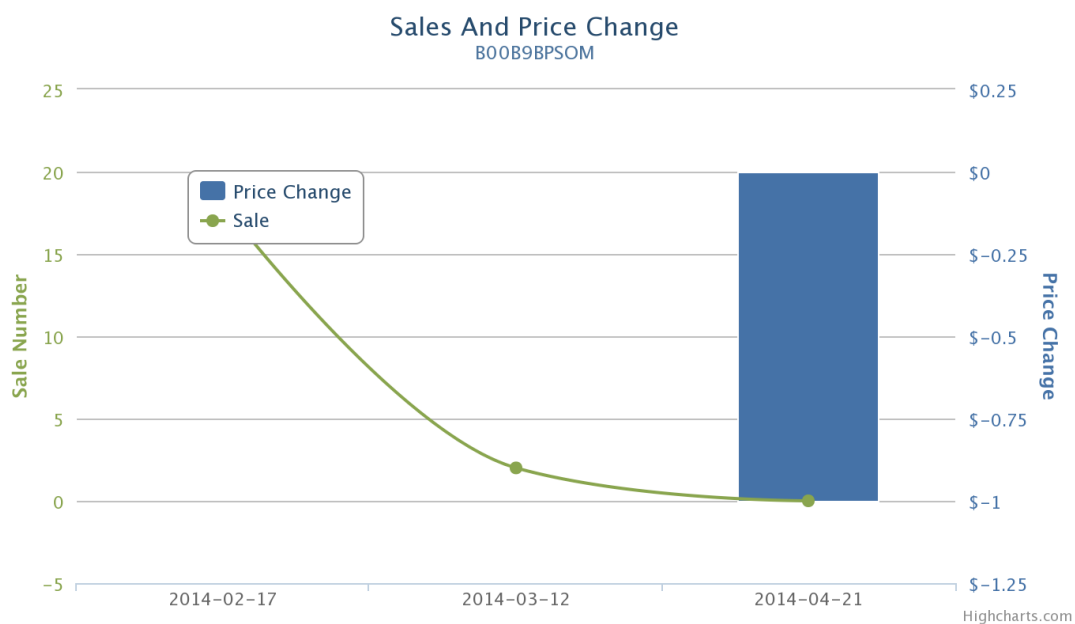
这里以 ASIN 为 B00DZ3RU2A 的商品为例，分析评论长度与评分的关系。



从此图中可以看出，长度较大的评论出现在高水平的评价上，而 1 星、2 星的评价，评语都短短带过。高评价的短评论也存在一定的聚集。由于此商品的评论信息不够多，无法准确判断两者之间的关系，因此此猜测有待今后进一步验证。

6. 分析价格走势与评论数量增量的关系

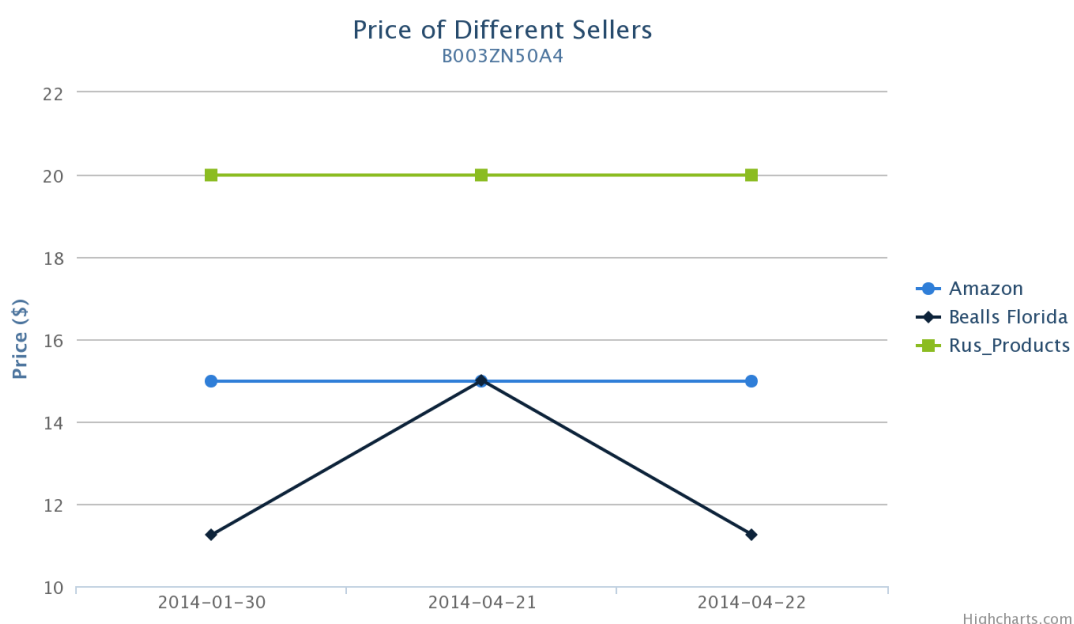
这里以 ASIN 为 B00B9BPSOM 的商品为例，分析价格走势与评论数量增量的关系。



从此图中可以看出，价格下降并没有产生更多的订单。这与我们的预期并不相符，原因可能与选择的商品销售数据不完全足够有关。此关系有待今后获取到一批准评论数量更多的商品后再进行研究。

7. 分析不同卖家出售的同一商品的价格走势

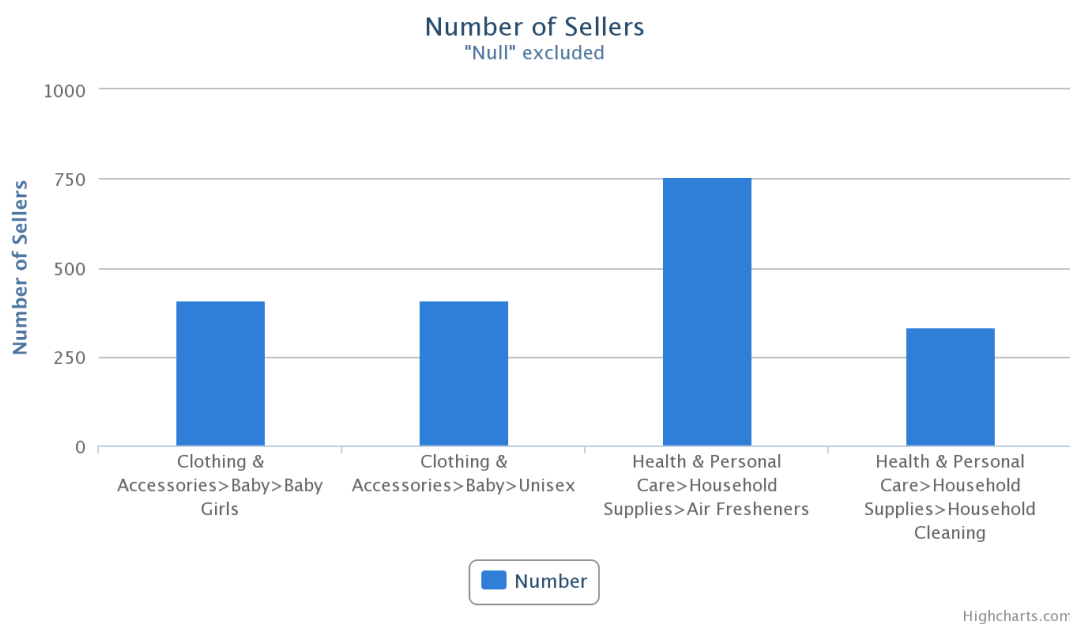
这里以 ASIN 为 B003ZN50A4 的商品为例，分析不同卖家出售的同一商品的价格走势。



从图中可以看出，第三方商家 Rus_Products 的价格始终最高，而 Bealls Florida 的商品价格始终处于低水平。对于 Rus_Products 而言，此图表明该商家需要尽快下调价格以在商品销售中获得先机。

8. 出售或某一类别商品的商家数量

这里以之前选定的四个分类为例，比较出售某一类商品的商家数量。



从图中可以看出，出售空气清新剂的商家在四个分类中最多，达到750户。从之前的准备工作中，我们得知空气清新剂分类内的商品数量比童装要少。基于这个前提，我们可以认为空气清新剂是亚马逊上销售的热门商品类目。

四、多余的话

四个分类是怎么选出来的？其实我们就是随便找了四个分类出来.....

上面的八张图表，这里实在不敢说挑出了最具代表性的商品来绘图，毕竟四个分类里面也是上千种商品，实在没法一个一个做人工的判断。所以，我们就直接选了几个销售商户多的、评论数量多的、价格抓取记录比较密集的商品出来绘图了，选择的标准并不是十分合适。

不过话说回来，就算没有把四个分类的所有商品全部看一遍，我们至少还是查看了一百多个商品的信息的。上面说到的一些功能，很多都因为爬取的数据量，尤其是价格和销售的数据量不够而没能展示出足够好的效果。助教们能不能有选择地选取一两件商品进行高频率的数据爬取，以获得更密集的数据呢？相信这样一来，可以让很多现在看来效果不佳的分析功能发挥作用。