网易严选商品数据分析

摘要

根据已有的2018年网易严选商品属性和商品评价的数据，本文主要通过Python进行一系列统计分析，主要内容为：对商品属性进行描述性统计分析，包括商品的类别、价格、销量、星级分布情况；对商品名称进行文本分析，用词云进行可视化；探究商品的价格和星级对销量的影响。

一、背景介绍

网易严选是网易旗下自营生活家居品牌，深度贯彻“好的生活，没那么贵”的品牌理念。网易严选秉承网易一贯的严谨态度，深入世界各地，与全球最优质的供应商进行合作，从挖掘消费需求出发，按需订制，全程参与把控工艺生产环节，为消费者提供好价格、好商品和好服务的优质体验。根据已经抓取的网易严选商品属性和商品评价数据，如何挖掘出其中的价值？对于买家来说，网易严选中的哪些商品销量高、口碑好、价格实惠？对于卖家来说，销量高、口碑好的商品都有哪些特点？商品的销量和星级、价格等是否有显著的相关性。这些都是下面将要探讨的问题。

二、数据分析

（一）关于价格、销量和星级评分的描述性统计

首先，通过Python的pandas模块读取包含商品属性的Excel文件，通过groupby函数按照商品类别（category）进行汇总，得到下表。

表2-1 商品类别及其商品数量表

类别	文体	居家	饮食	婴童	特色区	鞋包配饰	服装	洗护	电器	餐厨
数量	547	393	348	292	285	285	253	175	169	169

可以看到，网易严选将商品分为了十类：文体、居家、饮食、婴童、特色区、鞋包配饰、服装、洗护、电器、餐厨。在数据集中，文体、居家、饮食相关的商品数量最多，其他类型的商品数量相对平均。

我们对价格（price）进行描述性统计分析，得到表2-2和2-3。

表2-2 各种类别商品价格的描述性统计表1

	均值	标准差	偏度	峰度
文体	140.28	229.33	7.01	66.55
居家	692.53	1445.69	3.64	16.77
饮食	109.18	319.40	8.59	87.76
婴童	141.44	185.44	4.69	29.87
特色区	304.23	962.99	8.83	92.44
鞋包配饰	225.30	172.81	2.10	172.81
服装	187.50	148.53	2.94	9.54
洗护	75.43	78.55	2.56	8.79
电器	283.40	500.11	5.36	38.59
餐厨	134.65	165.51	3.08	12.91

从整体上来看，居家类型的商品价格最高，且远高于其他类型的商品。其次是特色区、电器和鞋包配饰类的商品。而洗护类型的商品价格最低。同时，各种类型的商品的标准差都比较大，说明同种类型的商品都有不同价格层次的商品。而从偏度和峰度则可明显看出，各种类型的商品价格有非常显著的右偏特征且极端值较多。表 2-3 的四分位数也证实了上述特征。

表 2-3 各种类别商品价格的描述性统计表 2

	最小值	第一四分位数	中位数	第三四分位数	最大值
文体	3.9	49	89	149	2999
居家	9.9	49	119	459	11999
饮食	6	18	28	90.65	3968
婴童	4.9	58.75	99	140.75	1799
特色区	9.9	49	108	206	11999
鞋包配饰	9	99	209	299	1399
服装	19.9	89	169	229	1299
洗护	4.9	24.95	49.9	99	520
电器	12.9	59	129	269	4680
餐厨	6.9	39	79	159	1199

除了价格外，销量也是买家关心的指标。我们对销量进行同样的描述性统计分析。

表 2-4 各种类别商品销量的描述性统计表

	均值	标准差	偏度	峰度
文体	571.93	2428.45	7.46	63.92
居家	2760.09	5324.81	5.68	48.14
饮食	5465.61	8317.10	3.30	14.31
婴童	1376.65	4751.23	13.39	205.80
特色区	1883.63	3410.89	4.38	25.88
鞋包配饰	5256.01	13024.85	9.72	125.32
服装	4263.06	8603.77	5.20	34.79
洗护	6090.34	11528.33	3.62	14.48
电器	3927.17	8904.60	7.10	66.72
餐厨	4743.96	9414.25	4.99	32.71

从整体上来看，洗护、饮食、鞋包配饰等类型的商品销量较高，文体类型的商品销量最低。同时，各种类型的商品的标准差都非常大，说明同种类型的商品销量差异很大。而从偏度和峰度则可明显看出，各种类型的商品价格有非常显著的右偏特征且极端值很多。

我们还知道每个商品的全部评价数据，其中星级评分是买家和卖家最为关注的。我们对星级评分（star）也进行描述性统计分析。值得注意的是，星级评分的范围为 1-5，而我们在读取数据的时候发现，在 star 这一列中出现了大量的 0。查看后面文字评价的内容，我们发现大多数都是积极正面的评价，因此我们可以推断出出现这种情况的原因是有部分买家忘记评分，导致最后的星级评分显示为 0。我们在进行分析时，需要把星级评分为 0 的数据当做缺失值，即删掉星级评分为 0 的数据。同时，我们注意到部分商品的销量较少，为了使商品的星级评价更加全面，我们过滤掉了有效星级评价小于 100 的商品。最终得到下表。

表 2-5 过滤筛选后商品星级评分的描述性统计表

	总数	均值	标准差	最小值	第一四分位数	中位数	第三四分位数	最大值
星级	2021	4.90	0.07	4.27	4.88	4.92	4.95	5

从表中可以看到，总体来看商品的星级评分非常高，即使是最小值星级评分也有 4.27，这也和网易严选的理念“从挖掘消费需求出发，按需订制，全程参与把控工艺生产环节，为消费者提供好价格、好商品和好服务的优质体验”相吻合。同时，商品间星级评分的差异也很小。

（二）关于商品名称的文本分析

为了更加详细的了解商品的特点，我们对商品名称（prodName）进行文本分析。使用 jieba 分词，并用 wordcloud 绘制词云图进行可视化。

通过前期的分词测试，我们发现 jieba 分词时会保留一些我们不感兴趣的、没有太多实际意义的词语和一些数字以及特殊符号，同时部分词语如“抱枕”“手机壳”“黑猪肉”以及“守望先锋”“网易云音乐”这种专有名词会被拆分而失去原来的意思。因此，我们在进行分词前需要进行相关的预处理。首先，我们根据最初分词的结果，添加了一些新词，如“抱枕”、“守望先锋”等，防止后续在分词时被拆分。然后，我们找出一些停用词（包括一些数字和特殊符号，如“0”、“（”、“，”等），对商品名称进行过滤。因为在文本分析中我们要统计每个词语出现的个数，所以 为了准确性我们对过滤后的数据中的每个列表的元素去重，即每个标题被分割后的词语唯一。

我们挑选出现次数最多的 50 个词，得到如下的词云图。

图 2-1 商品名称词云图

从词云图中，我们可以直观地感受到商品名称含有词语的特点。从类型来看，服饰类（服装、鞋包配饰）商品中含有许多相同的词，如男式、女式、儿童、婴儿，而数量最多的服饰有 T 恤、短袖、套装、内裤、长袖等，材质多为棉质（纯棉、全棉）和牛皮，类型有经典、休闲、运动、复古等。从产地来源来看，有较多商品来自日本和韩国。从商品特点来看，经典、收藏版、天然、舒适、多功能、便携、简约、安全、柔软是商品名称中出现较多的词。此外，阴阳师、魔兽、守望先锋等游戏类商品也有不少数量。

（三）价格与销量的关系

我们再来看看价格与销量之间的关系。通过绘制散点图进行可视化，如下图所示。

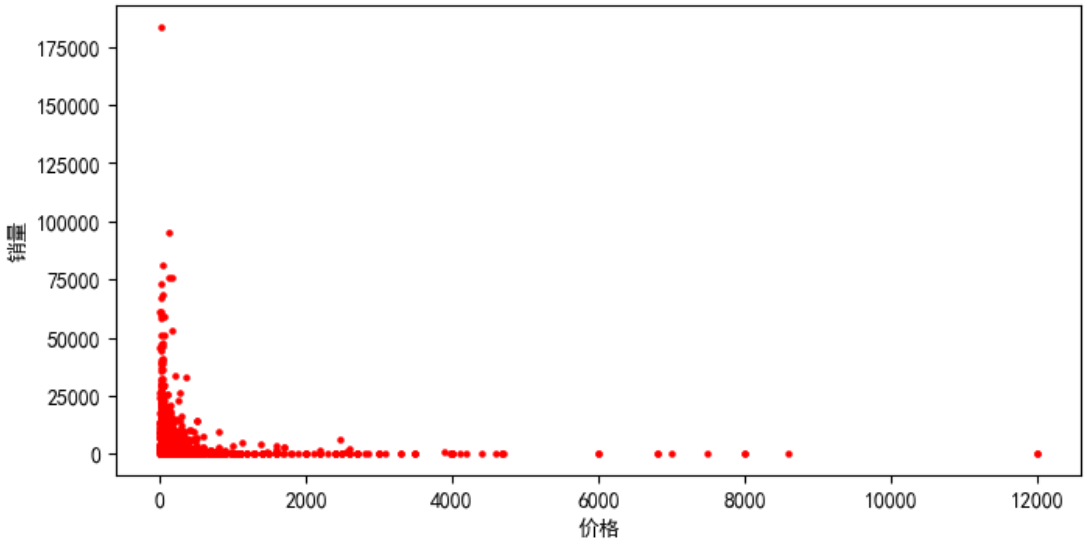

图 2-2 商品价格与销量的散点图

由图可知，总体趋势为随着商品价格增加其销量减少，商品价格对其销量影响较大；价格在500元以下的商品销量大多较高，且少数商品的销量冲的很高，价格大于500元后商品价格越高其销量越低。按照商品价格由高到低对所有商品进行排序，我们发现价格较高的商品大多数为家具和家用电器，如按摩椅、沙发、桌椅组合、燃气灶、电压力锅等。价格较低的商品大多数为文体和饮食类商品，这也和之前的描述性统计分析的结果一致。

相比买家更加关注销量，卖家则对销售额更加关注，即销量乘价格的值。通过商品属性文件，我们计算出每个商品的销售额，然后对其做关于价格的线性回归，结果如下图。

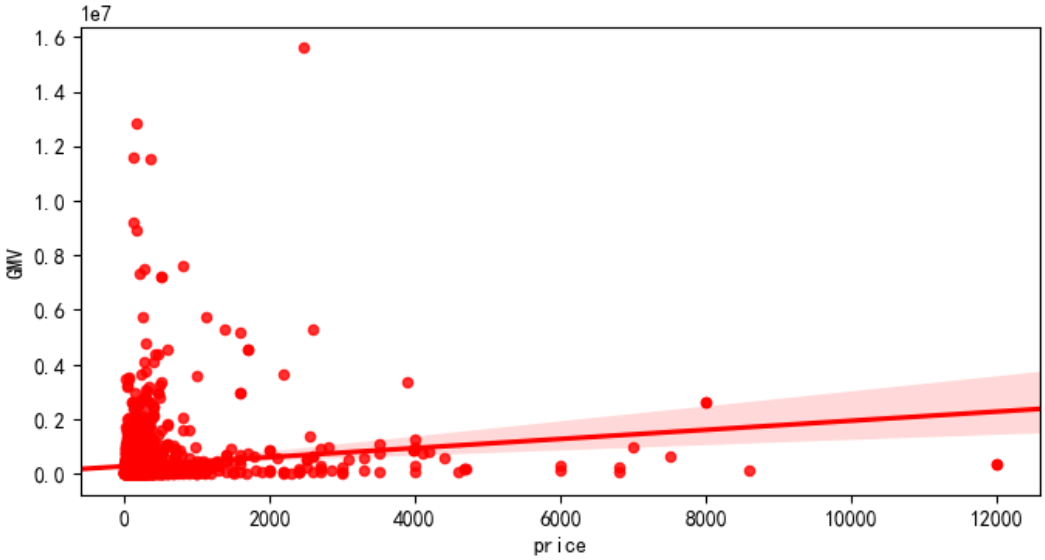

图 2-3 商品价格与销售额的散点图与线性回归模型

由线性回归的拟合线可以看出，总体趋势为商品销售额随着价格增长呈现一定的上升趋势；多数商品的价格偏低，销售额也偏低；销售额较高的商品中，价格在 2000 元以下的商品最多，少数商品价格在 2000-4000 元；价格超过 4000 元的商品销售额都不高，主要原因还是因为销量很少。

（四）星级与销量的关系

我们再看一下星级和销量之间的关系。通过绘制散点图进行可视化，如下图所示。

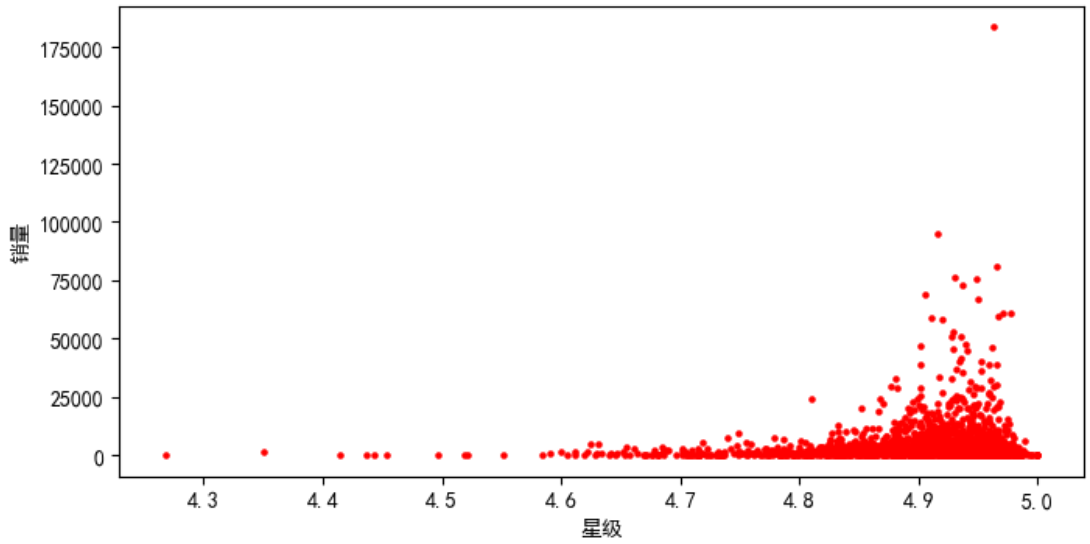

图 2-4 商品星级与销量的散点图

由图可知，总体趋势为随着商品星级评分增加其销量增加，商品价格对其销量影响较大；销量较高的商品平均星级评分大多在 4.9 以上，而平均星级评分低于 4.7 的商品销量都较少。

三、结论

根据前面的一系列分析，我们可以得到以下结论：

（1）商品类型

网易严选将商品分为了十类：文体、居家、饮食、婴童、特色区、鞋包配饰、服装、洗护、电器、餐厨。其中，文体、居家、饮食相关的商品数量最多，其他类型的商品数量相对平均。

（2）商品价格

从整体上来看，居家类型的商品价格最高，且远高于其他类型的商品。其次是特色区、电器和鞋包配饰类的商品。而洗护类型的商品价格最低。同种类型的商品都有不同价格层次

的商品，各种类型的商品价格有非常显著的右偏特征且极端值较多。

（3）商品销量

从整体上来看，洗护、饮食、鞋包配饰等类型的商品销量较高，文体类型的商品销量最低。同时，同种类型的商品销量差异很大，各种类型的商品价格有非常显著的右偏特征且极端值很多。

（4）商品星级

总体来看商品的星级评分非常高，即使是最小值星级评分也有 4.27，这也和网易严选的理念“从挖掘消费需求出发，按需订制，全程参与把控工艺生产环节，为消费者提供好价格、好商品和好服务的优质体验”相吻合。同时，商品间星级评分的差异也很小。

（5）商品名称

从类型来看，服饰类（服装、鞋包配饰）商品中含有许多相同的词，如男式、女式、儿童、婴儿，而数量最多的服饰有 T 恤、短袖、套装、内裤、长袖等，材质多为棉质（纯棉、全棉）和牛皮，类型有经典、休闲、运动、复古等。从产地来源来看，有较多商品来自日本和韩国。从商品特点来看，经典、收藏版、天然、舒适、多功能、便携、简约、安全、柔软是商品名称中出现较多的词。此外，阴阳师、魔兽、守望先锋等游戏类商品也有不少数量。

（6）商品价格与销量、销售额的关系

商品价格与销量的总体趋势为随着商品价格增加其销量减少，商品价格对其销量影响较大；价格在 500 元以下的商品销量大多较高，且少数商品的销量冲的很高，价格大于 500 元后商品价格越高其销量越低。

商品价格与销售额的总体趋势为商品销售额随着价格增长呈现一定的上升趋势；多数商品的价格偏低，销售额也偏低；销售额较高的商品中，价格在 2000 元以下的商品最多，少数商品价格在 2000-4000 元；价格超过 4000 元的商品销售额都不高

（7）星级与销量的关系

总体趋势为随着商品星级评分增加其销量增加，商品价格对其销量影响较大；销量较高的商品平均星级评分大多在 4.9 以上，而平均星级评分低于 4.7 的商品销量都较少。