

商务数据分析大作业

2018011787 核 81 李灵

2021.12

目录

1 前言	3
1.1 分析对象	3
1.2 分析背景	3
1.3 分析目的	3
1.4 分析思路	3
2 数据清洗	3
2.1 数据清洗的原则和依据	3
2.2 数据清洗的过程和结果	3
2.2.1 缺失值	3
2.2.2 无用字段	4
2.2.3 重复字段	4
2.2.4 分类	4
3 统计分析	4
3.1 基本统计指标计算	4
3.1.1 各产品所获评论数	4
3.1.2 各用户评论数	4
3.1.3 用户评论词频较高的单词	6
3.2 统计分析结论	6
4 数据分析	7
4.1 数据分析逻辑	7
4.1.1 心智模型	7
4.1.2 对心智模型的解释	7
4.2 加工过程	7
4.2.1 产品分析	7
4.2.2 用户分析	8
4.3 数据可视化	8

4.3.1	不同评价词频分析	8
4.3.2	销量分析	11
5	结论	11
5.1	结果、规律、知识发现	11
5.2	分析结果的商业解释	12
6	建议	12
6.1	商务应用的前景	12
6.1.1	应用领域和方式	12
6.1.2	商业模式设计	12
6.2	伦理道德的思考	12

1 前言

1.1 分析对象

这个数据集 [1] 包括了亚马逊的各种精美食品的评论。这些数据跨越了 10 多年的时间，包括了从 1999 年 10 月至 2012 年 10 月的所有评论。评论包括产品和用户信息、评级和纯文本评论，还包括亚马逊所有其他类别的评论。

该数据集内共有 568,454 条评论，来自 256,059 个用户，评论对象为 74,258 个产品。其中，260 名用户有超过 50 条评论。

1.2 分析背景

商业数据的价值只有通过分析才能得到。当前众多互联网公司都依托于各自的平台对用户进行数据的收集，对这些数据进一步分析对公司未来的业务发展方向可能起到尤为重要的作用。因此需要依托数据分析工具对大量数据进行快速、准确的处理。

1.3 分析目的

通过对亚马逊的精美食品的评论数据的分析，挖掘出其中有价值的部分并加以处理获得新知识，从而对公司决策有所裨益。

通过这些评论数据获取整体商品的销售数据，为产品更新换代提供数据支持，为平台推荐商品提供数据支撑，进一步发掘了解客户评价的重要性。

1.4 分析思路

挖掘整体营收指标：包括展示的商品种类、客户数量、商品销量。

对销售量前几名的产品进行研究，查看其评分。

查看销量榜前几占总销量的比重。

查看各评分段的产品的销量。

2 数据清洗

2.1 数据清洗的原则和依据

数据清洗的原则主要包括：非空检验、非法值清洗、数据格式检验、记录数检验、重复检验。

主要依据是处理后是否能够得到标准的、干净的、连续的数据以供后续数据分析使用。

2.2 数据清洗的过程和结果

2.2.1 缺失值

使用代码或 Excel 寻找缺失值，并未发现。

2.2.2 无用字段

使用代码或 Excel 寻找乱码或空白字段，并未发现。

2.2.3 重复字段

使用代码或 Excel 寻找一模一样的字段，并未发现。

2.2.4 分类

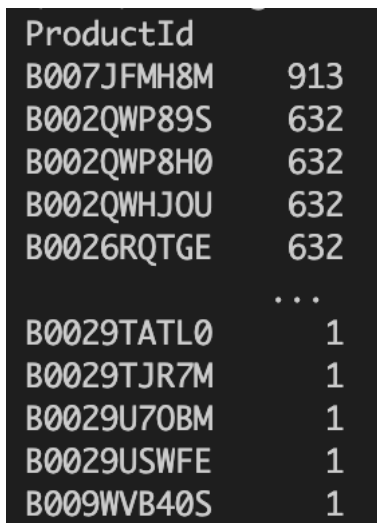
考虑到文本字段对评分分析影响不大，将数据集分为评分 (Reviews.csv) 和文本 (text.csv) 两个文件进行分析。

3 统计分析

3.1 基本统计指标计算

3.1.1 各产品所获评论数

通过代码对数据集进行分组计算，得到各产品所获评论数由大到小排序如下1。



ProductId	
B007JFMH8M	913
B002QWP89S	632
B002QWP8H0	632
B002QWHJOU	632
B0026RQTGE	632
...	
B0029TATL0	1
B0029TJR7M	1
B0029U70BM	1
B0029USWFE	1
B009WVB40S	1

图 1: 各产品所获评论数

取所获评论数前十的产品如2。将其转换为矩形图为3。

可以看到，在评论数前十的产品中，只有一个产品的评论量大于了 800，其余的都只超过了 600。对于这些评论数较多的产品，不论评价如何，如果不考虑销量造假的情况，那么其评论量也正反映了其购买量，这一数字显然也是很可观的。

3.1.2 各用户评论数

通过 amazon.py 程序对数据集进行计算，得到各用户所有的评论数由大到小排序如下4。

ProductId	
B007JFMH8M	913
B002QWP89S	632
B002QWP8H0	632
B002QWHJOU	632
B0026RQTGE	632
B003B300PA	623
B001E05Q64	567
B000VK8AVK	564
B007M8330Z	564
B0026KNQSA	564

图 2: 所获评论数前十的产品

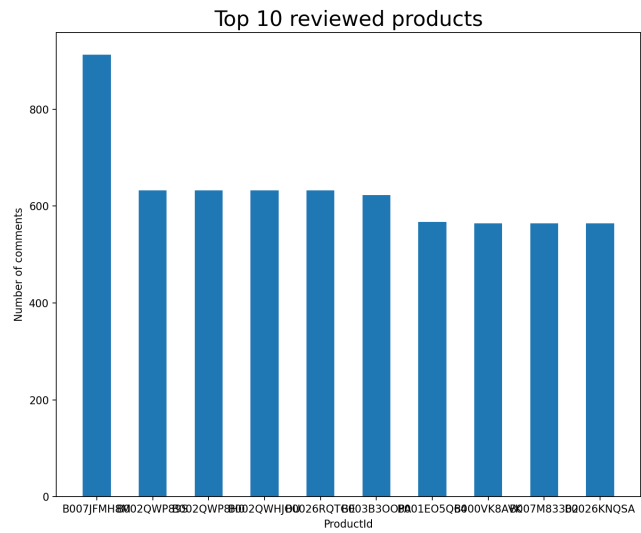


图 3: Top 10 reviewed products

UserId	
A30XHLG6DIBRW8	448
A1YUL9PCJR3JTY	421
AY12DBB0U420B	389
A281NPSIMI1C2R	365
A1Z54EM24Y40LL	256
...	
A2HROKQ00GA5AF	1
A2HROR28DMJV2W	1
A2HRR8C02Y20G8	1
A2HRSML93IK9TR	1
AZZZOVIBXHGDR	1

图 4: 各用户所有评论数

取所有评论数前十的用户 ID 如5。将其转换为矩形图为6。
可以看到，在评论数前十的用户中，有两位贡献了超过了 400 条评论，其余的也有 200 条左右。

UserId	
A30XHLG6DIBRW8	448
A1YUL9PCJR3JTY	421
AY12DBB0U420B	389
A281NPSIMI1C2R	365
A1Z54EM24Y40LL	256
A1TMAVN4CEM8U8	204
A2MUGFV2TDQ47K	201
A3TVZM3ZIXG8YW	199
A3PJZ8TU8FDQ1K	178
AQQLWCMRNDFGI	176

图 5: 所有评论数前十的用户 ID

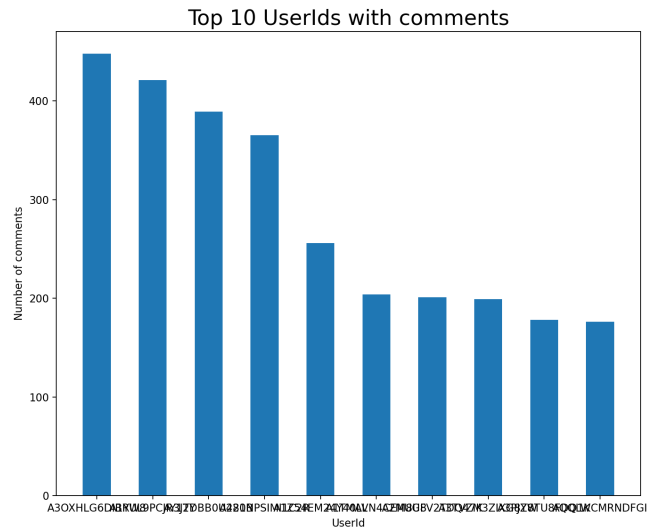


图 6: Top 10 UserIds with comments

3.1.3 用户评论词频较高的单词

对 text.csv 文件中的 score 项和 summary 项进行数据分析，使用 word.py 程序将各分数段对应的词频前 50 的单词存入 word.csv 文件以供后续分析。

3.2 统计分析结论

见后文。

	word	times	score	percent/comment
0	not	8019	1	0.153421
1	the	4679	1	0.089519
2	this	3070	1	0.058736
3	a	3039	1	0.058143
4	and	2821	1	0.053972
..
245	yum	4908	5	0.013516
246	but	4723	5	0.013007
247	are	4685	5	0.012902
248	not	4424	5	0.012183
249	them	4339	5	0.011949

图 7: Top 50 words occurring in each marking band

4 数据分析

4.1 数据分析逻辑

4.1.1 心智模型

由 csv 文件预览可知, 该数据集包括 ProductId、UserId、ProfileName、HelpfulnessNumerator、HelpfulnessDenominator、Score、Time、Summary、Text。从个人角度来说, HelpfulnessNumerator、HelpfulnessDenominator、Time 这三项是与数据本身关联度不大的内容, ProfileName 可能会涉及用户隐私, 因此均不予分析。

集中对 ProductId 和 UserId 进行分析, 着重分析其本身数量的关系和与 Score 之间的联系。对于 Summary 和 Text, 利用词频分析等工具分析其与 Score 之间的联系, 从而得到有价值的信息。

由于该数据集的评价数据也正是销售数据, 因此从企业角度来说也可以着重关注那些销量较高的产品, 对于评分略低但销量可观的产品的评价予以关注。同时, 对于平均评分高于 4 分的产品, 也可以关注其实时反馈, 对于偏差较大的评价及时反馈, 便于维护产品和企业口碑。

4.1.2 对心智模型的解释

如果是单纯作为一名数据分析人员, 我可能会更在意 Score 与其他项之间的关联; 但考虑到企业有通过分析数据获得利益的需求, 因此需要更在意分析能够带来利益的几项。

4.2 加工过程

进一步对评论超过 800 条的产品和超过 400 条评论的两位用户进行分析。

4.2.1 产品分析

B007JFMH8M: 在该产品的所有评价中, 最高分为 5 分, 最低分为 1 分, 平均分为 4.58 分, 相比于 5 分的满分来说该评价较好。由词频统计可得, 用户对该产品做出较好评价 (分数在 3 分以上) 的次数也远远高于做出较低评价的次数。

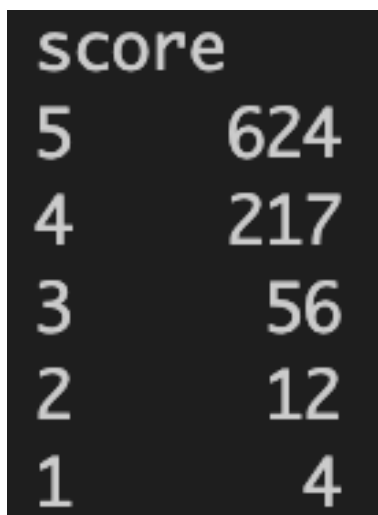


图 8: AB007JFMH8M 产品的用户评价分布

4.2.2 用户分析

(1) A3OXHLG6DIBRW8: 在该用户的所有评价中, 最高分为 5 分, 最低分为 2 分, 平均分为 4.138 分, 相比于 5 分的满分来说较高。由词频统计可得, 该用户对产品做出较好评价 (分数在 3 分以上) 的次数远远高于做出较低评价的次数。

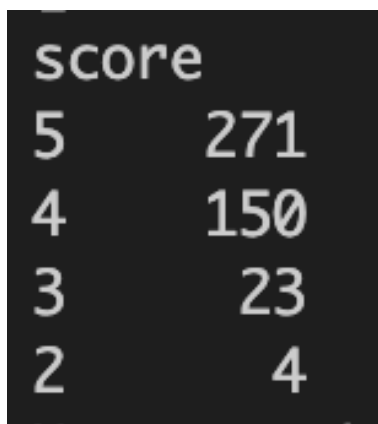


图 9: A3OXHLG6DIBRW8 用户的评价分布

(2) A1YUL9PCJR3JTY: 在该用户的所有评价中, 最高分为 5 分, 最低分为 2 分, 平均分为 4.494 分, 相比于 5 分的满分来说仍然较高。由词频统计可得, 该用户对产品做出较好评价 (分数在 3 分以上) 的次数也远远高于做出较低评价的次数。

4.3 数据可视化

4.3.1 不同评价词频分析

对用户评价中出现次数较高的词语做词云, 根据分数评价从 1-5 依次如下。

score	
5	240
4	150
3	30
2	1

4.3.2 销量分析

销量前十的产品及销量如下图16。

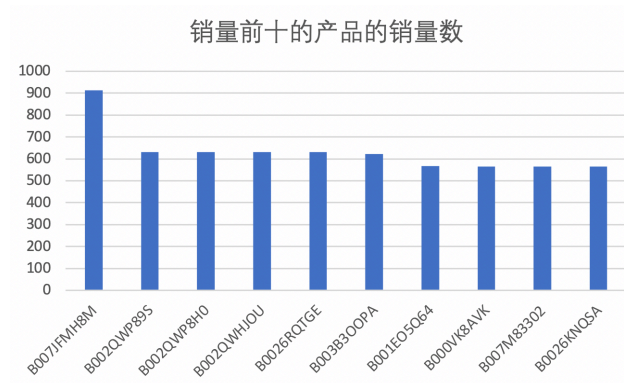


图 16: 销量前十的产品及其销量

销量前 100 的产品销量之和占总销量之比如下图17。



图 17: 销量前 100 的产品销量之和占总销量之比

5 结论

5.1 结果、规律、知识发现

可以看到，评价较差（评价分数小于等于 3）时，出现评论最高的词语都是 not，因此我们将评价分数 3 也归入评价较差里是情有可原的。同时，评价较好（评价分数大于 3）时，good 和 great 出现频

率最高，进一步表现了用户评价时对产品的喜爱。因此，光凭借评分判断用户的反馈也许并不够全面，可以结合词频进行进一步分析，对情感特别强烈的词语赋予更高的权重，可能会获得不一样的结果。

5.2 分析结果的商业解释

评论次数较多的产品通常是人们购买意愿较强的、购买次数较多的，因此其评分很显然不会低。评论次数较多的用户也是一样，购买意愿较强，很可能会多次购买自己喜欢的产品进行评价，予以剔除后可能分析效果会更好一些。

6 建议

6.1 商务应用的前景

6.1.1 应用领域和方式

对于评论数较多的产品和好评数较多的产品，都可以使用算法进行推荐或是与供应商商量收取一定的广告费用。

对于评论数量较多、评论质量较为优质的用户予以一定的奖励，对于特别优质的评论进行推广。

6.1.2 商业模式设计

通过分析销量得到销量较高的产品，便于公司进行下一步的推广；对于评分略低但销量可观的产品的评价予以关注，对公司进行反馈；对于相较于平均分偏差较大的评价反馈给公司，便于公司及时行动。

同时，对于词频的分析可以帮助企业进一步完善评分机制，从而得到真正意义上的“口碑产品”。

6.2 伦理道德的思考

本数据集收集的数据多为用户自行评论，对其进行收集其实会涉及到用户隐私问题，也许这已经被包含在用户隐私协议中，但我们不得而知。尽力隐去个体的信息可能也是分析需要注意的地方，当然本次分析中并没有需要精准到个体的地方，因此在数据分析时均略去了用户名称而以后台的用户 ID 代替。

另外，如何应用这些分析结果也是涉及伦理的一大问题。在“大数据杀熟”和“知识茧房”日益泛滥的当下，我们不得不反思是否过于依赖这些数据和大数据精准推送了。

参考文献

[1] Amazon-fine-food-reviews. <https://www.kaggle.com/snap/amazon-fine-food-reviews>.