Relationship Between Variables

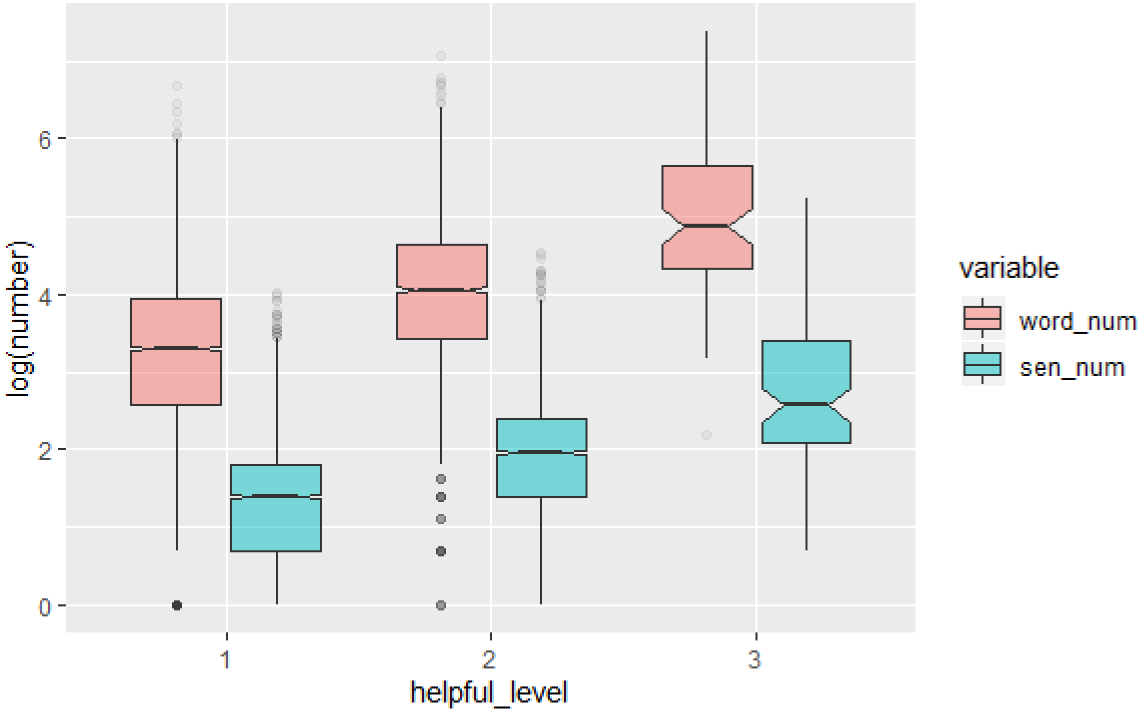
The following graphs are about hair dryer dataset. Pacifier and microwave share the same traits in this part.

Preprocessing

In order to find the relationship between star rating, reviews and helpfulness rating, it's necessary to make preliminary modifications to the original data, and observe the distribution of the three variables respectively. Since the majority of samples have $helpful_voting$ values of 0, and a few of them have large values, $helpful_voting$ is divided into three levels, the frequency and the corresponding relationship with $helpful_voting$ are as follows:

（插入表格helpful_level）.

$helpful_level$ & $word_num$,$sen_num$



- The higher the $helpful_level$, the more words there are in the comments and the more sentences there are. It is reasonable to assume that because more detailed reviews describr the product informatively to other buyers, thus they receive more $helpful_votings$.
- The height of boxes at different levels is almost the same. The comment characteristics of $helpful_level$ are relatively stable.

$helpful_level$ & $star_rating$

chi-square test

- In order to test whether there is a correlation between these two discrete variables, the chi-square test should be used. The null hypothesis is that the two discrete variables are independent of each other. The test results are as follows:

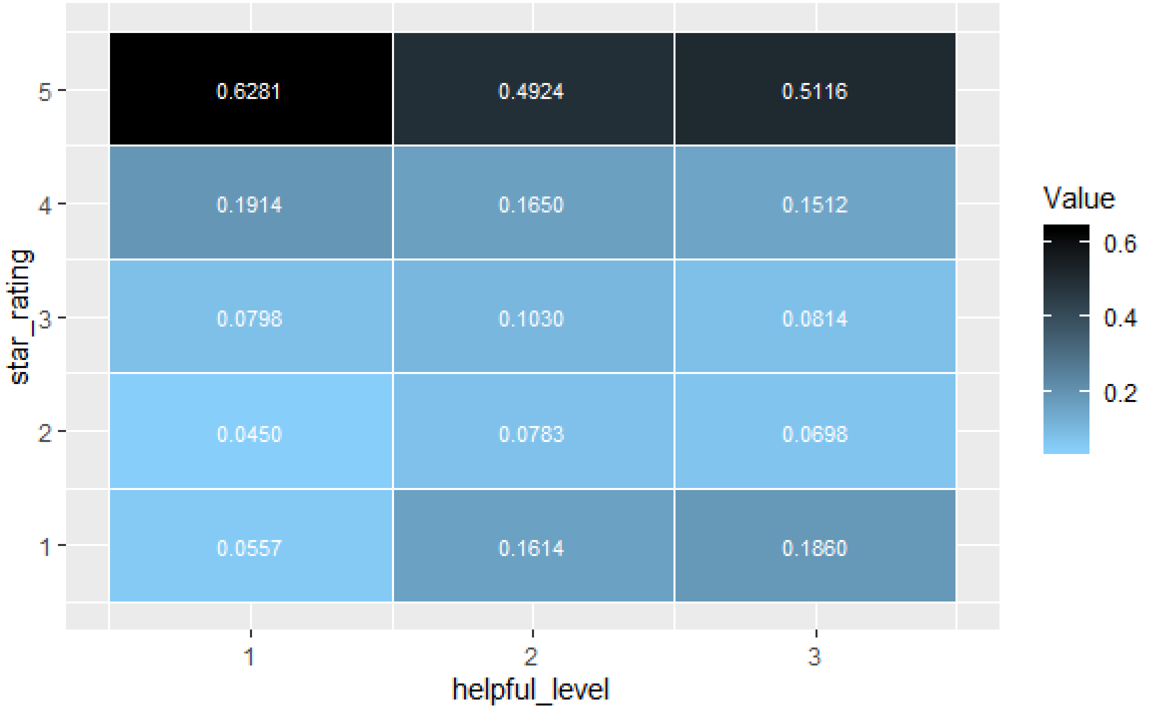
  (插入表格chi_square_test)

According to this p-value, we reject the null hypothesis. $helpful_level$ & $star_rating$ not independent.

Visualization

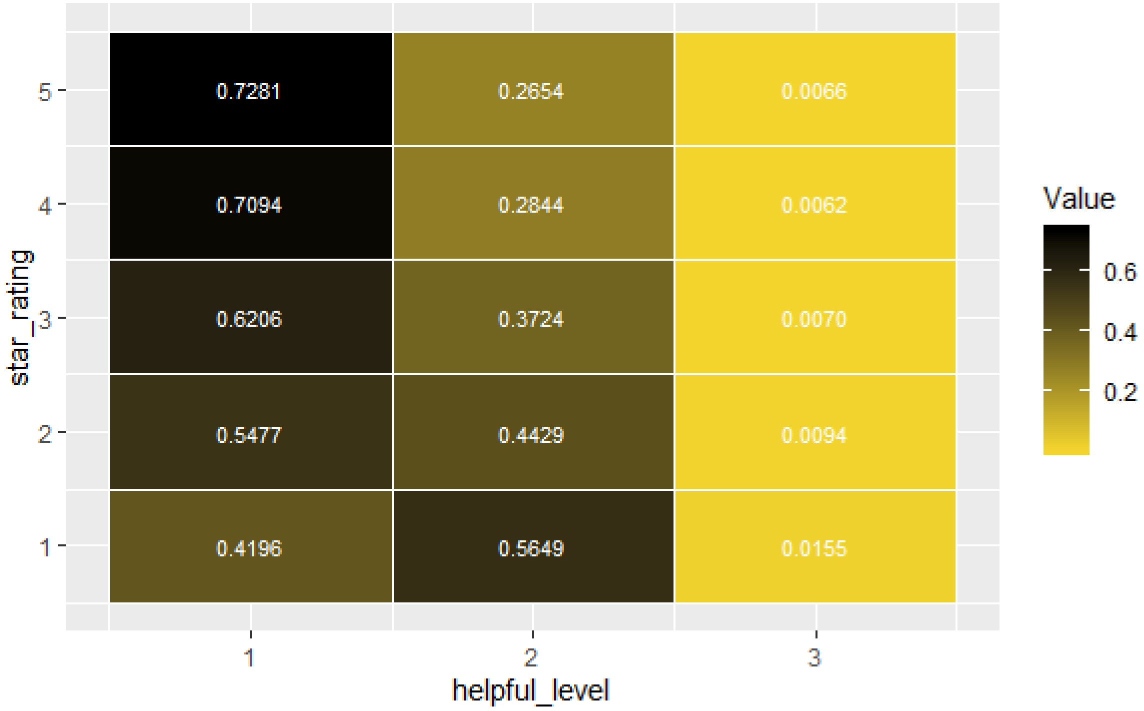
使用HeatMap 观察 helpful_level 和 star_rating 之间的具体关系

1.



- 上图中的数据表示某星级评论的评论对应 helpful_level 中的占比，**纵向**之和为1.
- 结论
    - 首先观察到，HeatMap最上面一行颜色最深，表示无论是哪个helpful_level，五星评论总是最多的，四星评价次之，总体来看，消费者对吹风机的评价还是比较高的。
    - 评论数目最少的星级为二星级，helpful_level 为1的评价一星评价略少于三星，而 helpful_level 为2或者3的评论一星评价远多于二星，三星评价。这可能由于在低星评价中会出现一些关于产品缺点的细致描写，这些信息在其他评论中是少见的，而这些信息恰好是消费者需要的信息，故其helpful_level也比较高

2.



- 上图中的数据表示某helpful_level评论的评论在对应star_rating中的占比，**横向**之和为1.
- 结论：
  - 对于星级评价为2,3,4,5的评价，表中最左侧一列颜色最深，表示无论是哪个star_rating，大多数 help_level 均为1的，其次是2，最后是2，这是由于这三者频数差距较大，helpful_level为1的样本频数最大。
  - 对于一星评价，占比最多的helpful_level是2，虽然helpful_level 为1 的频数几乎是 helpful_level 为 2 的两倍。

小结

由上关系和模式可知，helpful_level为2/3的评论本身语言更加丰富，单词数目和句子个数都更多，而低星评价也更可能出现在这两类当中，Sunshine公司应当注意helpful_level较高的评论的内容的信息，方便对其自身的产品进行改进。