Question 2 d

思路

根据前 30 天的 star_rating 对后 7 天的 reviews 六个特征分别进行回归，即，用户看到前一个月评价的 star_rating 加权均值是否对其评论特征产生影响。

- 自变量：
  - 该时间段普通用户评级均值（不是vine 也不是verified）
  - 该时间段vine用户的评级均值
  - 该时间段verified_purchase评级均值
- 因变量：
  - 基于评论内容信息的特征：CF
    - 评论长度 word_num
    - 句子数量：sen_num
    - 平均句子长度：word_per_sen
  - 基于商品特征信息的评论特征：PF
    - 商品特征数量：用名词来衡量：n_num
    - 特征情感词数量：用形容词来衡量：adj_num
    - 行为词数量：用动词来衡量：vb_num

数据筛选

为了降低关系中的不稳定因素，我们只选取在该段时间内评论数目大于2并且在其对应时间区间评论条数也大于2的样本进行回归分析

变量相关性的描述

- 2*3的图矩阵（吹风机数据）
- 结论：
  - 观察图片可以看到，如果该用户看到的前一个月的打分均较高或者均较低，那么该用户评论的长度，句子数量，商品特征数量，特征情感词数量，行为词数量均较多，相反，如果该用户看到的前一个月的打分居中，比如大多为三星四星评价，那么该用户的评论上述五个指标会更低。
  - 平均句子长度的规律与其他五个特征不同，用户若接连看到低评或者接连看到高评，那么他的评论的平均句子长度较短，相反，会较高。

回归分析

- 假设上述关系如果存在，则为线性关系。
- 使用逐步回归的方法（verified_rate均没有通过t检验，故被排除，最后只剩下常数项，common_rate还有vine_rate）
- 为了验证上述特征在数值上存在相关关系，下面分别用三个自变量对因变量进行回归，观察得到的线性模型是否显著，如果显著，则认为显著相关，否则，不可认为相关
- 这里插入回归系数以及 p 值表格

- 结论：
  - 观察上表 p 值，我们可以有95%的把握认为这三个自变量与句子长度，句子个数，平均每个句子单词数，行为词数量这四个变量有关，但在该置信区间上，与商品特征数量，特征情感词数量这两个变量没有直接关系。
  - 观察变量的显著性检验结果，常数项高度显著，common_rate比较显著，vine_rate不太显著，verified_purchase被模型排除，可知用户的评价虽然与前一个月的评级有关，但是不会因此产生过大的波动，并且当用户阅读评级时，写该评级的人是否是 vine 只会产生一点影响，而写该评级的人是否 verified_purchase 几乎没有影响