Intro to Data Science HW 1

Ben Green

Contents

February 22, 2015

Question 1

1a.

Feature Name	Feature Type
`producer`	nominal
`release_to_review_time`	interval
`used_real_name`	binary
`verified_purchase`	binary
`rating`	ordinal
`helpfulness`	ordinal
`number_of_votes`	ratio
`length_of_review_text`	ratio

1b.

The mode of `producer` is Apple, with 4480 entries.

1c.

5077/9585, or 53% of reviewers used their real name and had a verified purchase.

1d.

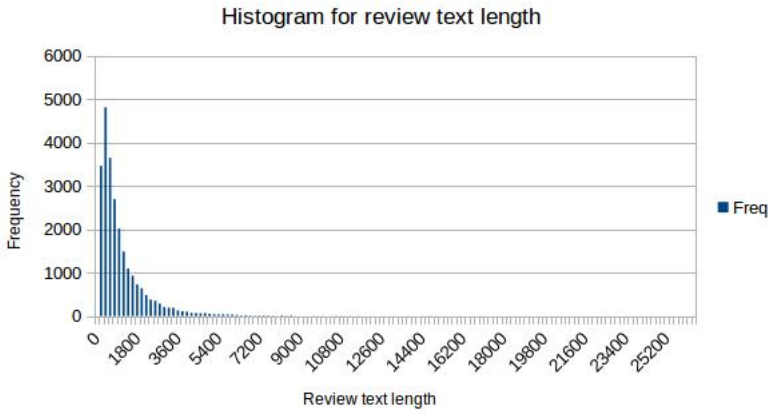
5077/13989, or 36% of reviews with a verified purchase had a reviewer who used their real name.

1e.

Measure	**Value**
min	-537
q1	74
median	144
max	11686
interquartile	215

These numbers can be displayed conveniently in a boxplot.

1f.



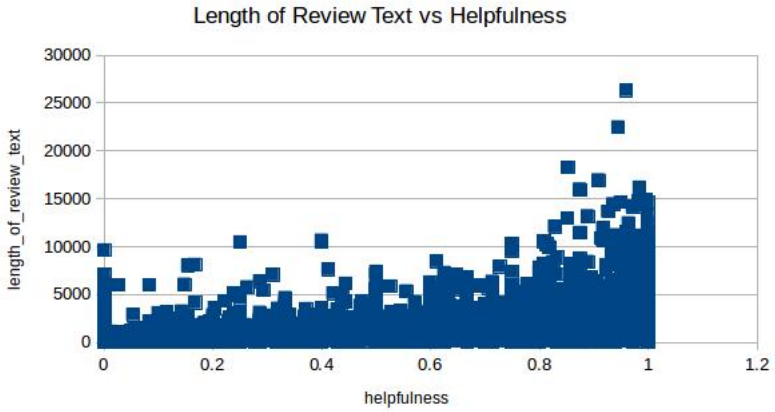
1g.

Yes, the distribution of `length_of_review_text` is heavily skewed towards the shorter end. There are outliers of 26332, 1, and 1?

1h.

The Pearson correlation value between `length_of_review_text` and `helpfulness` is approximately .25. This indicates that there is a slight positive correlation between these two variables.

1i.



Question 2