

Is there seasonality in star ratings?

Data Analytics Assignment 3 | Joshua Cooper | Student ID: 16098824 | Group C

23 March 2017

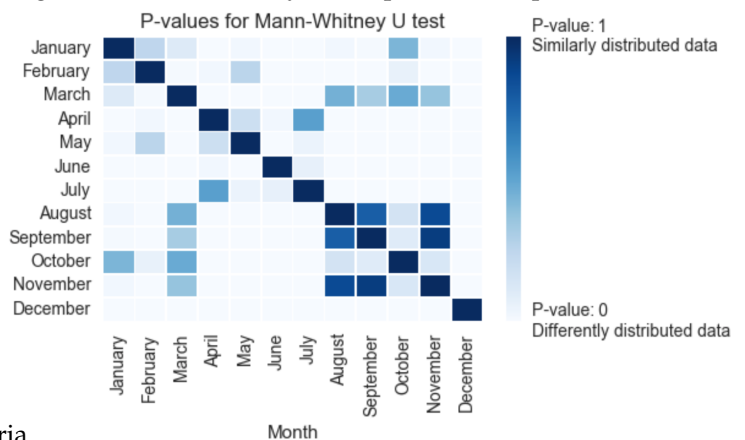
Is there seasonality in star ratings?

Seasonality seems to be present in the data, as indicated by a range of statistical tests.

Testing for seasonality using monthly distributions of ratings

Seasonality was first tested by grouping all star ratings in the dataset into the month in which they were submitted, forming twelve samples which, in aggregate, formed the entire dataset. These twelve monthly samples were tested to see whether they could plausibly arise from the same distribution. After determining that the distribution of star ratings is not of Gaussian form, it was observed that an ANOVA test would be inappropriate for our data. Instead, the non-parametric Kruskal-Wallis test was used. The results of this test for the entire dataset convincingly indicated, with an effectively zero p-value, that at least two of the monthly samples had to be generated by different distributions. More granularity was then sought to determine whether there was one outlier month, or whether most pairs of months seemed to be generated by different distributions. The Mann-Whitney U test is analogous to a Kruskal Wallis test when examining two samples, so the Mann-Whitney U test was performed on each pair of monthly samples. The results are shown opposite, and reveal that for most (48 out of 66) pairs of monthly samples, we reject the null hypothesis at the 5% significance level that the pair of samples arise from the same distribution, implying abundant seasonality. Given the heterogeneity in the dataset, analysis was also conducted that tested whether seasonality existed in the most commonly reviewed cities and business types. Eighteen of the most commonly reviewed thirty cities tested positively for seasonality, using Kruskal Wallis at the 5% significance level, but twenty-nine of thirty most commonly reviewed business types tested positively under the same criteria.

Figure 1: Mann-Whitney U test p-values for pairs of months



Testing for seasonality using weather data

Weather data was additionally used to explain temporal variety in the data. Time series weather data for the ten most commonly reviewed cities were correlated with ratings on the corresponding dates in those cities. It was observed that for four of the ten cities, there was a statistically significant correlation between temperature and star rating, at a 5% significance level, using both Spearman and Pearson calculations. Three of these correlations were positive, while one was negative.

Conclusions

According to the statistical tests used, there seems to be a significant amount of seasonality in the data. However, the key for business owners is to consider seasonality in the context of their own business. This might mean filtering by both business type and location. Once a stakeholder arrives at the conclusion that there is seasonality, a seasonal adjustment may be made, which would allow the business owner to filter away a certain amount of "noise".

However, there are limitations in this approach. It may be the case that the majority of reviews are not written on the same day that the reviewer experienced the services of the business. There may also be other factors related to weather, such as humidity and precipitation, that correlate with ratings. Furthermore, the relationship between rating and temperature may be non-linear, and non-linearity has the potential to render correlation ineffective. For those wishing to extend this analysis, there are additional data that can be used to further detect or explain seasonality, such as tourism data.