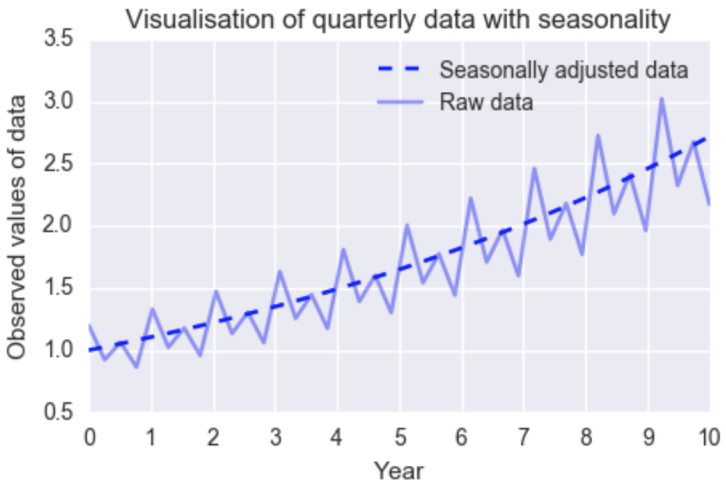
Is there seasonality in star ratings?

Data Analytics Assignment 2 | Joshua Cooper | Student ID: 16098824 | Group C

23 February 2017

1 Abstract

Data, when collected over time, often has seasonality. This is especially prevalent in economics or in data related to sales. For instance, an ice cream vendor can expect sales to be much higher in the summer than in the winter. It is important for the vendor to be commercially aware of this fact for several reasons. Firstly, the vendor should buy more supplies for summer in preparation for increased demand. However, it is also important for the vendor to know that, as sales fall as Autumn approaches, this trend does not imply a worsening in the quality of his/her product. In other words, businesses should simultaneously exploit seasonality and be careful not to attribute trends caused by seasonal effects to business decisions. The plot shown opposite (made using toy data) shows quarterly seasonality.



2 Application to the Yelp dataset

Businesses on Yelp are likely to be interested in change over time in their user ratings. A deterioration in average star rating might prompt changes in business strategy, and a rise in ratings might make management think a particular strategy, deployed at a lucky time, is successful. However, people's social behaviour may be influenced by seasonal factors (such as warmer weather in summer or a more jovial atmosphere during the lead up to Christmas), and this may feed into a change in overall ratings. Therefore, businesses need to be aware of seasonality in their ratings, so that they can make informed business decisions grounded in sound reasoning.

3 Experiment

Our null hypothesis is that there is no seasonality in the data, and we will perform statistical experiments using established methods to determine whether the null hypothesis holds. The alternative hypothesis is that there is seasonality present in the data. There is a variety of statistical tests that can be deployed to address this problem. The reviews can be grouped by different periods in time, and then we can run an ANOVA test to see whether these time-bucketed data seem to be generated from the same distribution. However, if the data being measured are not normally distributed, we should instead use the Kruskal-Wallis test, a non-parametric test for comparing samples from multiple groups. We can also test correlation of star ratings over time with seasonal data, such as weather and tourism data. If there is a lack of seasonality among the entire data set, this may simply be a result of the heterogeneity (with respect to location, business type, etc.) observed within the data, and more exploration will be required to see whether seasonality exists at a more granular level. For example, we might find seasonality in particular industries or cities. We have approximately ten years of data with which we can conduct this experiment.